

Second Edition

INTRODUCTORY MODERN ALGEBRA

A Historical Approach

SAUL STAHL

$$x_1x_2 + x_3x_4$$

$$x + y\sqrt{-2}$$

$$x^{p^v} - x$$

WILEY

Introductory Modern Algebra

Introductory Modern Algebra A Historical Approach

Second Edition

Saul Stahl

Department of Mathematics
University of Kansas
Lawrence, KS

WILEY

Copyright © 2013 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.

Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permission>.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representation or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print, however, may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Stahl, Saul.

Introductory modern algebra : a historical approach / Saul Stahl, Department of Mathematics, University of Kansas. — Second edition.

p. cm.

Includes bibliographical references and index.

ISBN 978-0-470-87616-9 (cloth)

1. Algebra, Abstract. I. Title.

QA162.S73 2013

512'.02—dc23

2013018928

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Contents

Preface	ix
1 The Early History	1
1.1 The Breakthrough	1
2 Complex Numbers	9
2.1 Rational Functions of Complex Numbers	9
2.2 Complex Roots	17
2.3 Solvability by Radicals I	23
2.4 Ruler-and-Compass Constructibility	26
2.5 Orders of Roots of Unity	36
2.6 The Existence of Complex Numbers*	38
3 Solutions of Equations	45
3.1 The Cubic Formula	45
3.2 Solvability by Radicals II	49
3.3 Other Types of Solutions*	50
4 Modular Arithmetic	57
4.1 Modular Addition, Subtraction, and Multiplication	57
4.2 The Euclidean Algorithm and Modular Inverses	62
4.3 Radicals in Modular Arithmetic*	70
4.4 The Fundamental Theorem of Arithmetic*	71
5 The Binomial Theorem and Modular Powers	75
5.1 The Binomial Theorem	75
5.2 Fermat's Theorem and Modular Exponents	85
5.3 The Multinomial Theorem*	90
5.4 The Euler φ -Function*	93

* Optional

6	Polynomials over a Field	99
6.1	Fields and Their Polynomials	99
6.2	The Factorization of Polynomials	107
6.3	The Euclidean Algorithm for Polynomials	113
6.4	Elementary Symmetric Polynomials*	120
6.5	Lagrange's Solution of the Quartic Equation*	125
7	Galois Fields	131
7.1	Galois's Construction of His Fields	131
7.2	The Galois Polynomial	139
7.3	The Primitive Element Theorem	144
7.4	On the Variety of Galois Fields*	147
8	Permutations	155
8.1	Permuting the Variables of a Function I	155
8.2	Permutations	158
8.3	Permuting the Variables of a Function II	166
8.4	The Parity of a Permutation	169
9	Groups	183
9.1	Permutation Groups	183
9.2	Abstract Groups	192
9.3	Isomorphisms of Groups and Orders of Elements	199
9.4	Subgroups and Their Orders	206
9.5	Cyclic Groups and Subgroups	215
9.6	Cayley's Theorem	218
10	Quotient Groups and Their Uses	225
10.1	Quotient Groups	225
10.2	Group Homomorphisms	234
10.3	The Rigorous Construction of Fields	240
10.4	Galois Groups and Resolvability of Equations	253
11	Topics in Elementary Group Theory	261
11.1	The Direct Product of Groups	261
11.2	More Classifications	265

12	Number Theory	273
12.1	Pythagorean Triples	273
12.2	Sums of Two Squares	278
12.3	Quadratic Reciprocity	285
12.4	The Gaussian Integers	294
12.5	Eulerian Integers and Others	304
12.6	What Is the Essence of Primality?	310
13	The Arithmetic of Ideals	317
13.1	Preliminaries	317
13.2	Integers of a Quadratic Field	319
13.3	Ideals	322
13.4	Cancellation of Ideals	337
13.5	Norms of Ideals	341
13.6	Prime Ideals and Unique Factorization	343
13.7	Constructing Prime Ideals	347
14	Abstract Rings	355
14.1	Rings	355
14.2	Ideals	358
14.3	Domains	361
14.4	Quotients of Rings	367
A	Excerpts: Al-Khwarizmi	377
B	Excerpts: Cardano	383
C	Excerpts: Abel	389
D	Excerpts: Galois	395
E	Excerpts: Cayley	401
F	Mathematical Induction	405

G	Logic, Predicates, Sets, and Functions	413
G.1	Truth Tables	413
G.2	Modeling Implication	415
G.3	Predicates and Their Negation	418
G.4	Two Applications	419
G.5	Sets	421
G.6	Functions	422
	Biographies	427
	Bibliography	431
	Solutions to Selected Exercises	433
	Index	444
	Notation	448

Preface

IT IS COMMON KNOWLEDGE amongst mathematicians that much of modern algebra has its roots in the issue of solvability of equations by radicals. The purpose of this text is to provide the undergraduate mathematics majors and the prospective high school mathematics teachers with a one-semester introduction to modern algebra that keeps this relationship in view at all times.

Most modern algebra texts employ an axiomatic strategy that begins with abstract groups and ends with fields, ignoring the issue of solvability of equations by radicals. By contrast, we follow the paper trail from the Renaissance solution of the cubic equation to Galois's description of his ideas. In the process, all the important concepts are encountered, each in a well-motivated manner.

One year of calculus provides all the information required for the comprehension of all the topics in this text, which has many distinguishing features:

Historical development. Students would prefer to know the real reasons that underlie the creation of the mathematical structures they encounter. They also enjoy being placed in direct contact with the works of the prime movers of mathematics. This text tries to bring them as close to the source as possible.

Finite groups and fields are rooted in some specific investigations of Lagrange, Gauss, Cauchy, Abel, and Galois regarding the solvability of equations by radicals. This text makes these connections explicit. Gauss's proof of the constructibility of the regular 17-sided polygon is incorporated into the development, and the argument given is merely a paraphrase of that which appears in the *Disquisitiones*. Similarly, the proof of Theorem 8.10 is just a reorganization of that given by Abel in his paper on the quintic equation. The construction of Galois fields is accomplished in the form of a commentary on the opening pages of Galois's paper *On the Theory of Numbers* which are quoted verbatim in the text. Several important documents are also included as appendices. A considerable amount of historical discussion is integrated into the development of the subject matter.

Cohesive organization. The historical development of the material allows for very little flexibility. Each chapter elucidates some of the preceding material and motivates ideas

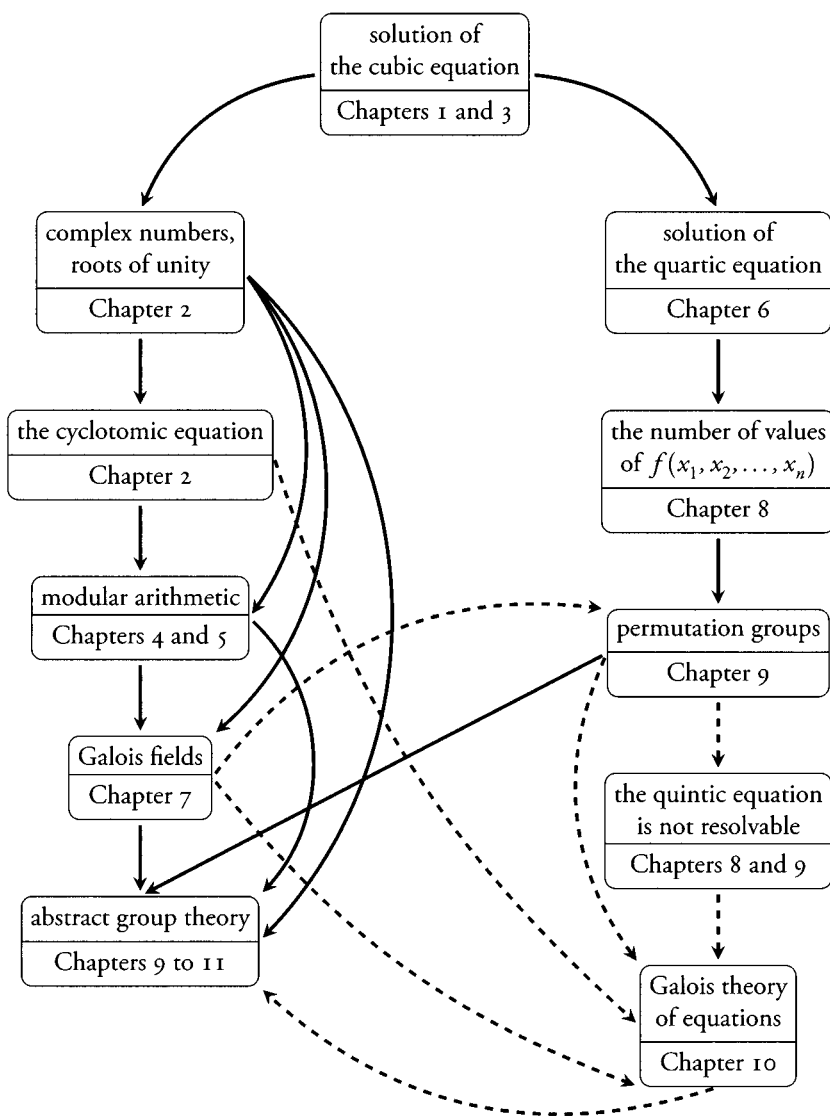


Figure 0.1 The genesis of the theory of finite groups.

that come later. The advantage of this approach is the same as that of good motivation in general: it aids comprehension by providing the students with a framework in which to

fit the various concepts they encounter. A one semester course can be constructed on the basis of Sections 1.1, 2.1–5, 3.1–2, 4.1–2, 5.1–2, 6.1–3, 7.1–3, 8.1–4, 9.1–5, & 10.1.

Figure 0.1 illustrates the author's perception of the evolution of abstract group theory (ignoring all the geometric and much of the number-theoretic contributions). The number in the right of each box denotes the chapter in which this topic is discussed. Solid arrows correspond to connections that are treated in some depth whereas those that are displayed by dashed arrows are touched on only informally.

Chapters 1 to 3 are dedicated to the formalization of the notion of solvability by radicals. Gauss's proof of the constructibility of the regular 17-sided polygon is the capstone theorem of this part of the course. Field theory is developed in Chapters 4 to 7. The Primitive Element Theorem of Section 7.3 serves as a watershed: it unifies many of the important concepts that precede it and motivates the notion of cyclicity that comes later. Group theory is developed in Chapters 8 to 10. This begins with an explanation of the relevance of permutations to solvability by radicals, goes on to the discussion of permutation groups and abstract groups, and concludes with the description of quotient groups. Chapter 11 is meant to acquaint the students with some of the standard tools of elementary group theory.

Exercises. Each section is followed by its own set of exercises. These range from the routine to the challenging. Each chapter has an additional set of easy review exercises added to remind the students of the chapter's main points. There are over 1,000 of these end-of-section and chapter review exercises. The answers to selected odd exercises appear at the end of the book. Most chapters are also accompanied by a collection of supplementary computer and/or mathematical projects. Some of the latter involve open questions.

Additional pedagogy. Each chapter begins with an introduction and concludes with a summary. The purposes of both the introduction and the summary are to provide the student with an overview of the chapter, and sometimes to comment on its relationship to the previous chapters. The examples are integrated into the exposition and they are highlighted by a notation in the margin. Each chapter's new terms are listed, together with the pages on which they are defined, following that chapter's summary.

Instructor's manual. An instructor's manual is available. It contains the answers to all the end of section and chapter review exercises. Some suggested homework assignments and tests are also included.

Acknowledgments

First and foremost I wish to acknowledge the substantial contributions made by Fred Galvin who rooted out several inaccuracies in the original development, improved and/or corrected many of the proofs, both in the text and the manual, suggested new exercises, and used the manuscript in his class. Thanks are also due to Todd Eisworth, Andy Magid and Phil Montgomery who also class tested the manuscript and made valuable suggestions as well as to my colleague Paul J. McCarthy who was kind enough to lend me both an ear and his algebraic expertise. It remains to gratefully acknowledge the efforts of Jessica Downey, Steve Quigley, Rosalyn Farkas, and Lisa Van Horn of John Wiley & Sons on behalf of this book.

June 1996

Preface to the Second Edition

Surprisingly, it turned out that the historical approach could be used to teach ring theory as well. The point of departure is the Theorem of Pythagoras, viewed as a diophantine equation. Chapter 12 begins there and goes on to Fermat's characterization of primes that are the sum of two square integers. From there we go on to quadratic reciprocity and the Gaussian integers. The question of Gaussian primes is natural and some attention is given to variant number systems with radicals $\sqrt{-2}$ or $\sqrt{-3}$. The chapter ends with a discussion of Kummer's decision to redefine the notion of primality.

Quadratic fields, quadratic integers, and ideals are defined and the arithmetic of ideals is explored in Chapter 13. It is shown that the arithmetic of ideals does possess the unique factorization property. Finally, Chapter 14 discusses rings and ideals in the abstract manner of today.

The author's understanding of the low level algebraic number theory in Chapter 13 comes from reading one of Keith Conrad's many expository monographs. The solutions to the selected exercises in Chapters 13 and 14 were derived by Grant Serio and are included with his permission. Katie Ballentine, Annika Denkert, and Mark Hunacek debugged portions of the manuscript, which was expertly typeset by Lon Mitchell.

June 2013

Saul Stahl
Lawrence, Kansas

Chapter 1



THE EARLY HISTORY

THIS CHAPTER CONTAINS an informal account of the early history of the issue of solvability of equations of degrees one, two, and three in a single unknown. The formulas that provide the solutions lead in a natural way to the discussion of the origins of complex numbers. We also take this opportunity to review some well-known information about the quadratic equation.

1.1 The Breakthrough

There is a general agreement among historians of mathematics that modern mathematics came into being in the mid sixteenth century when the combined efforts of the Italian mathematicians Scipione del Ferro, Niccolò Tartaglia, and Gerolamo Cardano produced a formula for the solution of cubic equations. For the first time ever west European mathematicians succeeded in cracking a problem whose solution eluded the best mathematical minds of antiquity. Archimedes, one of the greatest mathematicians, scientists, and engineers of all times, had solved some cubic equations in terms of the intersections of a suitable parabola and hyperbola. Omar Khayyam, one of the most prominent of the Arab mathematicians and poets, also expended much effort on his geometrical solutions of special cases of the cubic equation but could not find the general formula. However, the significance of this accomplishment of the Renaissance mathematicians is not limited to the difficulty of the problem that was solved. We shall try to show how the issues raised by this solution eventually led to the creation of modern algebra and the discovery of mathematical landscapes that were undreamt of, even by such imaginative investigators as Archimedes and Khayyam.

The interest in algebraic equations goes back to the beginnings of written history. The *Rhind Mathematical Papyrus*, found in Egypt circa 1856 is a copy of a list of mathematical problems compiled some time during the second half of the nineteenth century BCE, or

possibly even earlier. The twenty-fourth of these problems reads: “A quantity and its $1/7$ added become 19. What is the quantity?” In other words, what is the solution to the equation

$$x + \frac{x}{7} = 19?$$

The method employed by the scribe has come to be known as the *method of false position*. He replaces the unknown by 7 and observes that

$$7 + \frac{7}{7} = 8.$$

From this he concludes that the correct answer is obtained upon multiplying the first guess of 7 by $19/8$:

$$x = 7 \cdot \frac{19}{8} = \frac{133}{8}.$$

Interestingly enough, the scribe does double check his solution by substituting it into the original problem and verifying that

$$\frac{133}{8} + \frac{133/8}{7} = 19.$$

We will not discuss the merits and limitations of the method of false position except to note that the idea of obtaining a correct solution to an equation by starting out with a possibly false guess and then modifying that guess has been refined into powerful techniques for finding numerical solutions, one of which will be described in Section 3.3. We do, however, wish to point out that the general *first-degree equation* is today defined as

$$ax + b = 0, \quad a \neq 0,$$

and that the rules of algebra yield

$$x = -\frac{b}{a}$$

as its unique solution.

The Mesopotamian mathematicians of that time could solve much more intricate equations, and had in fact already developed techniques for solving what we nowadays call quadratic equations. These techniques employed the geometrical method of “completing the square.” The Greeks, Indians, and Arabs all were aware of this method, having either derived them independently or perhaps learnt them from their predecessors and/or neighbors. In the ninth century the Persian mathematician al-Khwarizmi (عَبْدَ اللَّهِ مُحَمَّدُ بْنُ)

المؤسى الخوارزمي wrote the book *Hisab al-jabr w'al-muqa-balah* (الكتاب المختصر في حساب الجبر والمقابلة) in which he carefully explained a compendium of algebraic techniques learnt from several past civilizations. The clarity of his exposition won both him and his book immortality in that the portion *al-jabr* of the title evolved into the word *algebra*, and the author's name is the source of the word *algorithm*. An excerpt from this book expounding the solution to the quadratic equation

$$x^2 + 10x = 39$$

appears in Appendix A. The modern solution of the quadratic also relies on the completion of the square. The general *quadratic equation* has the form

$$ax^2 + bx + c = 0, \quad a \neq 0, \quad (1.1)$$

and its solutions are found by first factoring out the coefficient a and then completing the rest to a perfect square. Thus, we first divide Equation 1.1 through by a to obtain the equation

$$x^2 + \frac{b}{a}x + \frac{c}{a} = 0. \quad (1.2)$$

The left side of Equation 1.2 is then transformed to a near perfect square:

$$\begin{aligned} x^2 + \frac{b}{a}x + \frac{c}{a} &= \left(x + \frac{b}{2a}\right)^2 - \frac{b^2}{4a^2} + \frac{c}{a} \\ &= \left(x + \frac{b}{2a}\right)^2 - \frac{b^2 - 4ac}{4a^2}. \end{aligned}$$

The original quadratic equation has thus been transformed to

$$\left(x + \frac{b}{2a}\right)^2 - \frac{b^2 - 4ac}{4a^2} = 0$$

or

$$\left(x + \frac{b}{2a}\right)^2 = \frac{b^2 - 4ac}{4a^2} \quad \text{or} \quad x + \frac{b}{2a} = \frac{\pm\sqrt{b^2 - 4ac}}{2a}.$$

Hence the general quadratic equation, Equation 1.1, has the two solutions

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \quad \text{and} \quad x_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}.$$

It is clear that if a , b , and c are real numbers, then these two solutions are real and distinct if $b^2 - 4ac > 0$, they are real and identical if $b^2 - 4ac = 0$, and they are imaginary and distinct if $b^2 - 4ac < 0$. Another important fact to bear in mind (Exercises 1.1.5 and 1.1.6) is that

$$x_1 + x_2 = -\frac{b}{a} \quad \text{and} \quad x_1 x_2 = \frac{c}{a},$$

from which it follows that it is easy to construct a quadratic equation whose roots are prespecified. As we will have several occasions to refer to these identities later, they are stated as a proposition whose proof is relegated to Exercise 1.1.14.

Proposition 1.3 For any two numbers r and s the quadratic equation

$$x^2 - (r + s)x + rs = 0$$

has r and s as its roots.

It is reasonable at this point to raise the ante and ask for a formula that will yield the solution of the general *cubic equation*

$$ax^3 + bx^2 + cx + d = 0. \quad (1.4)$$

There are indications that the Mesopotamians already tried to systematize the search for solutions of cubic equations, and we know for a fact that the Greeks attempted the same. As was mentioned above, the final breakthrough did not occur until the middle of the sixteenth century when it was shown that a solution of the equation

$$x^3 + px + q = 0$$

is given by the expression

$$x = \sqrt[3]{-q/2 + \sqrt{q^2/4 + p^3/27}} - \sqrt[3]{q/2 + \sqrt{q^2/4 + p^3/27}}. \quad (1.5)$$

As we shall see later, very little additional work is required to pass from this formula on to a formula for the general cubic equation (Equation 1.4), and so Formula 1.5 can be considered as the crucial step, even though it does not yield the solution to the most general cubic equation.

In analogy with the ancient solutions of the quadratic, this solution was obtained by a geometrical process of completing the cube. Excerpts from Cardano's description of

the solution are contained in Appendix B. A modern derivation of this formula appears in Chapter 3, and we restrict ourselves here to the examination of some instructive applications of Formula 1.5. Surprisingly, this formula raises some very interesting questions.

Consider the cubic equation $x^3 - 1 = 0$. Here $p = 0$ and $q = -1$, and so Formula 1.5 yields

$$x = \sqrt[3]{1/2 + \sqrt{1/4 + 0}} - \sqrt[3]{-1/2 + \sqrt{1/4 + 0}} = \sqrt[3]{1/2 + 1/2} - \sqrt[3]{-1/2 + 1/2} = 1,$$

which is as it should be. However, for the equation $x^3 + 6x - 20 = 0$, which Cardano uses as an illustration in his *Ars Magna*, the same formula yields the solution

$$x = \sqrt[3]{10 + \sqrt{100 + 8}} - \sqrt[3]{-10 + \sqrt{100 + 8}} = \sqrt[3]{\sqrt{108} + 10} - \sqrt[3]{\sqrt{108} - 10}.$$

It can be easily verified with the aid of a calculator that the above solution agrees with 2 to at least eight decimal places, and the mathematical verification that the agreement is absolute is left to Exercise 1.1.1. Our purpose in presenting this example was to draw attention to the possibility that Formula 1.5 may present a correct solution in an unnecessarily complicated form. This obfuscation becomes much more disturbing in the case of the equation $x^3 - 15x - 4 = 0$, treated by Rafael Bombelli in his *Algebra* (1572). Formula 1.5 yields the solution

$$x = \sqrt[3]{2 + \sqrt{-121}} - \sqrt[3]{-2 + \sqrt{-121}}. \quad (1.6)$$

However, it is easily verified by inspection that $x = 4$ is also a solution of this cubic, and, since

$$x^3 - 15x - 4 = (x - 4)(x^2 + 4x + 1),$$

two more solutions of the original equation are obtained by solving the quadratic

$$x^2 + 4x + 1 = 0.$$

As the solutions of this quadratic are $-2 \pm \sqrt{3}$, we are faced with the question of which of the three numbers 4 or $-2 \pm \sqrt{3}$ is disguised as Expression 1.6. Moreover, this complicated expression involves square roots of negative numbers, in other words, imaginary quantities, whereas 4 and $-2 \pm \sqrt{3}$ are all real numbers. This apparent paradox was

resolved by Bombelli who simplified Expression 1.6 by setting

$$\sqrt[3]{2 \pm \sqrt{-121}} = a \pm b\sqrt{-1},$$

cubing both sides and deriving $a = 2$ and $b = 1$ from the resulting simultaneous equations. Rather than exhibit the details of his solution we simply point out that indeed

$$\begin{aligned} (2 + \sqrt{-1})^3 &= 2^3 + 3 \cdot 2^2 \sqrt{-1} + 3 \cdot 2 \cdot (\sqrt{-1})^2 + (\sqrt{-1})^3 \\ &= 8 + 12\sqrt{-1} - 6 - \sqrt{-1} \\ &= 2 + 11\sqrt{-1} = 2 + \sqrt{-121} \end{aligned}$$

and similarly

$$(-2 + \sqrt{-1})^3 = -2 + \sqrt{-121}.$$

Consequently,

$$\sqrt[3]{2 + \sqrt{-121}} - \sqrt[3]{-2 + \sqrt{-121}} = 2 + \sqrt{-1} - (-2 + \sqrt{-1}) = 4.$$

Thus, users of the cubic formula ignore the so-called *imaginary numbers* at their peril. Such prejudices come at the cost of losing some real solutions to real equations. This is further borne out by the innocent-looking equation $x^3 - 3x = 0$. Formula 1.5 yields the solution

$$x = \sqrt[3]{\sqrt{-1}} - \sqrt[3]{\sqrt{-1}},$$

and even if one is very skeptical about the existence of imaginary quantities it is very tempting to believe in them just long enough for the above radicals to cancel out and to yield the root $x = 0$, which we know to be correct.

The solution to the cubic equation is the context within which imaginary numbers were first discussed by mathematicians. Cardano toyed with them and then rejected them as useless. Bombelli gave them more credence, but it wasn't until about 200 years later that the work of Leonhard Euler, Pierre-Simon de Laplace, and later that of Carl Friedrich Gauss, Augustin-Louis Cauchy, and Niels Abel turned the complex number system, consisting of both the real and imaginary numbers, into an indispensable tool for mathematical researchers.

The Ferro-Tartaglia-Cardano Formula 1.5 suffers from a serious deficiency. This formula yields at most one solution for any cubic equation, even when such an equation

is known to have three distinct real roots, as is the case for $x^3 - x = 0$ whose roots are 0 and ± 1 . In view of the fact that the quadratic formula of Equation 1.1 does succeed in incorporating all the solutions into one expression it would not seem unreasonable to expect the same of the cubic counterpart. As we shall see in the next chapter, the *complex numbers* will enable us to find just such an expression.

Exercises 1.1

1. Prove that $\sqrt[3]{\sqrt{108} + 10} - \sqrt[3]{\sqrt{108} - 10} = 2$.
2. Prove that $\sqrt{28 - 10\sqrt{3}} - \sqrt{7 - 4\sqrt{3}} = 3$.
3. Solve the equation $3x^2 - 2x - 2 = 0$.
4. Solve the equation $x^4 - 3x^2 + 2 = 0$.

If r and s are the roots of the quadratic equation $ax^2 + bx + c = 0$, prove the identities in Exercises 1.1.5 to 1.1.7.

5. $r + s = -b/a$
7. $r^2 + s^2 = (b^2 - 2ac)/a^2$
6. $rs = c/a$

If r and s are the roots of the quadratic equation $ax^2 + bx + c = 0$, rewrite the expressions in Exercises 1.1.8 to 1.1.13 in terms of a , b , and c . Wherever necessary, you may assume that the denominators are not zero.

8. $1/r + 1/s$
10. $r^2s + rs^2$
12. $1/r^2 + 1/s^2$
9. $r^3 + s^3$
11. $(r - s)^2$
13. $1/r^2s + 1/rs^2$

14. Prove Proposition 1.3.
15. If r and s are the roots of the equation $x^2 + px + q = 0$, what is the quadratic equation whose roots are $r + s$ and rs ?
16. If $r, s \neq 0$ are the roots of the equation $x^2 + px + q = 0$, what is the quadratic equation whose roots are $1/r$ and $1/s$?
17. For what real values of α are the roots of the equation $x^2 + \alpha x + \alpha = 0$ real?
18. For what values of m will the equation $x^2 - 2x(1 + 3m) + (3 + 2m) = 0$ have equal roots?

Chapter Summary

This introductory chapter was used to briefly review the solutions of the first- and second-degree equations in a single unknown. The history of the solution of the cubic equation was also discussed and the relationship of this formula to the complex number system was examined.

Chapter Review Exercises

Mark the following true or false.

1. Every real number is the solution of some equation.
2. Every pair of real numbers is the solution set of some quadratic equation.
3. Every equation has at least one solution.

New Terms

cubic equation, 4

first-degree equation, 2

method of false position, 2

quadratic equation, 3

Chapter 2



COMPLEX NUMBERS

THROUGHOUT HISTORY, the introduction of new numbers has been greeted with considerable resistance on the part of mathematicians. Legend has it that the discoverer of irrational numbers was rewarded by being drowned by his fellow Greeks. Be that as it may, the fact is that these numbers have been tagged with the pejorative label of *irrational*, a word which, when used in nonmathematical contexts, has definite derogatory connotations. The same, of course, applies to the *negative* numbers. The *imaginary* numbers have been cursed with what is arguably the worst nomenclature in mathematics. Given the considerable difficulties that the average students face in learning the rigorous discipline of mathematics, can they be blamed for balking at having to contend with quantities that mathematicians themselves admit are imaginary?

The best way to overcome people's resistance to a new concept is to convince them of its utility. Accordingly, it will be shown that the widening of our field of operations to include the complex numbers greatly enhances the power of the Ferro-Tartaglia-Cardano cubic formula. Next, the complex numbers will be used to solve some ruler-and-compass construction problems of plane geometry. Only in this chapter's last section will the issue of the existence of the complex numbers be addressed.

2.1 Rational Functions of Complex Numbers

Just as was done by the mathematicians of the eighteenth and nineteenth centuries, we assume here the existence of a number i which has the property that $i^2 = -1$.

The rigorous proof of i 's existence is deferred to Section 2.6. In the meantime, the number i is to be treated just like a variable, with the sole additional stipulation that whenever i^2 occurs within an algebraic expression, it can be replaced by -1 . A *complex number* is an expression of the form $a + bi$ where a and b are any real numbers. When

$b = 0$ such a number is called an *imaginary number* and when $b = 0$ it is said to be *real*. These complex numbers can be added and subtracted as polynomials. Thus,

$$(5 - 3i) + (-2 + 5i) = 5 - 3i - 2 + 5i = 3 + 2i,$$

$$(5 - 3i) - (-2 + 5i) = 5 - 3i + 2 - 5i = 7 - 8i.$$

The multiplication of complex numbers also resembles that of polynomials, except that each occurrence of i^2 is replaced by -1 . Thus,

$$\begin{aligned}(5 - 3i)(-2 + 5i) &= -10 + 25i + 6i - 15i^2 \\ &= -10 + 31i - 15(-1) \\ &= -10 + 31i + 15 = 5 + 31i.\end{aligned}$$

The division of complex numbers mimics the well-known process of rationalizing denominators. Thus,

$$\frac{5 - 3i}{-2 + 5i} = \frac{5 - 3i}{-2 + 5i} \cdot \frac{-2 - 5i}{-2 - 5i} = \frac{-10 - 25i + 6i - 15}{(-2)^2 - (5i)^2} = \frac{-25 - 19i}{4 + 25} = \frac{-25}{29} - \frac{19}{29}i.$$

Surprisingly, all of these arithmetical operations can be given very interesting visual, or geometric, interpretations. To accomplish this, we represent each complex number $a + bi$ by the point (a, b) of the Cartesian plane. The point (a, b) is called the *Cartesian representation* of the complex number $a + bi$. Given two complex numbers $a + bi$ and $c + di$, let their Cartesian representations be $P = (a, b)$ and $Q = (c, d)$ (Figure 2.1). Their sum

$$(a + bi) + (c + di) = (a + c) + (b + d)i$$

is represented by the point $R = (a + c, b + d)$. However,

$$\text{slope of } PR = \frac{(b + d) - b}{(a + c) - a} = \frac{d}{c} = \text{slope of } OQ$$

and

$$\text{slope of } QR = \frac{(b + d) - d}{(a + c) - c} = \frac{b}{a} = \text{slope of } OP.$$

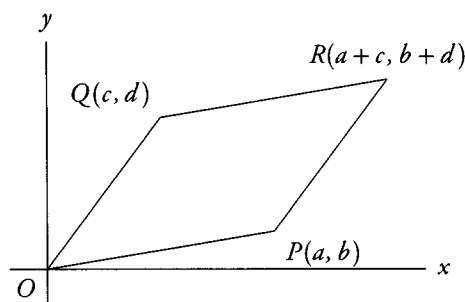


Figure 2.1 Complex addition

Consequently, $PR \parallel OQ$ and $QR \parallel OP$ and so $OPRQ$ is a parallelogram. Thus we see that the addition of complex numbers resembles that of vectors. These considerations are summarized as follows.

Proposition 2.1 Let O denote the origin of the Cartesian plane and let P and Q be the Cartesian representations of the complex numbers $a + bi$ and $c + di$, respectively. If the sum of the two complex numbers is represented by the point R , then the quadrilateral $OPRQ$ is a parallelogram.

To give the multiplication of complex numbers a visual interpretation, it is convenient to begin by establishing some conventions. In the sequel, the general complex number $a + bi$ will frequently be abbreviated as z . If either a or b is 0, it is omitted from $a + bi$. Thus, $3 + 0i = 3$ and $0 - 5i = -5i$.

Let $P = (a, b)$ be the Cartesian representation of the complex number $z = a + bi$ (Figure 2.2). The *modulus* of z , denoted by $|z|$, is the length of the line segment OP . In other words $|a + bi| = \sqrt{a^2 + b^2}$. Thus, for example,

$$|2 + 3i| = \sqrt{2^2 + 3^2} = 13,$$

$$|3 - 4i| = \sqrt{3^2 + (-4)^2} = 5,$$

$$|-2i| = \sqrt{0^2 + (-2)^2} = 2.$$

It is clear that the modulus of the real number $a + 0i$ is just its absolute value. Thus the modulus should be regarded as the extension of the notion of absolute value to the complex numbers. The *argument* of $z = a + bi$, denoted $\arg(z)$, is the counterclockwise angle from the positive x -axis to the ray OP where P is the Cartesian representation of z . As can be seen in Figure 2.3, the arguments of 3 , $1 + i$, $3i$, -2 , $-2 - 2i$, and

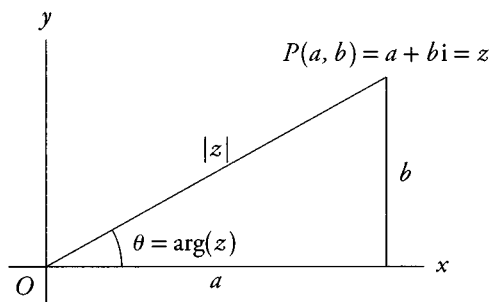


Figure 2.2 The argument and the modulus

$3 - 3i$ are $0, \pi/4, \pi/2, \pi, 5\pi/4$, and $7\pi/4$, respectively. For our purposes here it is convenient to identify angles whose measures differ by the full angle of 2π . Thus it will be convenient sometimes to regard 1 as having argument 2π or 4π rather than 0. The reasons for this will become clear after we have discussed the geometrical interpretation of the multiplication of complex numbers.

The argument θ of the general complex number $z = a + bi$ is easily computed (Figure 2.2) from the relation $\tan \theta = b/a$, but the quadrant in which z lies must be taken into account. Thus

$$\arg(1 + i) = \arctan \frac{1}{1} = \frac{\pi}{4},$$

whereas

$$\arg(-2 - 2i) = \pi + \arctan \frac{-2}{-2} = \frac{5\pi}{4}.$$

Observe that if the complex number $z = a + bi$ has argument θ , then, by Figure 2.2,

$$z = a + bi = |z| \left(\frac{a}{|z|} + \frac{b}{|z|}i \right) = |z|(\cos \theta + i \sin \theta).$$

We refer to $|z|(\cos \theta + i \sin \theta)$ as the *polar form* of z . For example, the complex numbers $1 + i$, 5 , i , and $-2i$ have polar forms $\sqrt{2}(\cos \pi/4 + i \sin \pi/4)$, $5(\cos 0 + i \sin 0)$, $\cos \pi/2 + i \sin \pi/2$, and $2(\cos 3\pi/2 + i \sin 3\pi/2)$, respectively. On the other hand, the number $3 + 4i$ has polar form $5(\cos \alpha + i \sin \alpha)$ where $\alpha = \arctan 4/3 \approx 53.13^\circ$.

Just as the addition of complex numbers has a geometrical interpretation in terms of Cartesian coordinates, so their multiplication can be easily visualized in terms of their polar forms. Let the complex numbers z and w be given in terms of their polar forms $z = |z|(\cos \theta + i \sin \theta)$ and $w = |w|(\cos \varphi + i \sin \varphi)$.

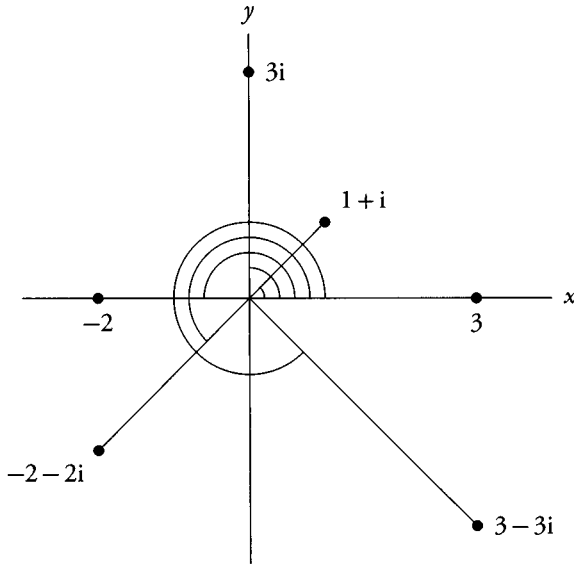


Figure 2.3 Some complex numbers

The trigonometric formulas for the functions of the sums of two angles yield

$$\begin{aligned}
 zw &= |z||w|(\cos \theta + i \sin \theta)(\cos \varphi + i \sin \varphi) \\
 &= |z||w|[(\cos \theta \cos \varphi - \sin \theta \sin \varphi) + i(\cos \theta \sin \varphi + \sin \theta \cos \varphi)] \\
 &= |z||w|[\cos(\theta + \varphi) + i \sin(\theta + \varphi)].
 \end{aligned}$$

Thus, the product zw has polar form $|z||w|[\cos(\theta + \varphi) + i \sin(\theta + \varphi)]$. Since it follows from Figure 2.2 that every complex number is completely determined by its argument and modulus, we conclude that $\arg(zw) = \theta + \varphi$ and $|zw| = |z||w|$. Hence, we have proved the following theorem.

Theorem 2.2 Let z and w be any two complex numbers. Then $\arg(zw) = \arg(z) + \arg(w)$ and $|zw| = |z||w|$.

If $z = i$ and $w = 1 + i$, then $zw = i(1 + i) = -1 + i$ and so

$$\arg(z) + \arg(w) = \frac{\pi}{2} + \frac{\pi}{4} = \frac{3\pi}{4} = \arg(zw)$$

and $|z||w| = 1 \cdot \sqrt{2} = \sqrt{2} = |zw|$.

Angles whose measures differ by integer multiples of 2π are considered to be identical. Thus, the number i has all the following as its arguments:

$$\dots, \frac{-7\pi}{2}, \frac{-3\pi}{2}, \frac{\pi}{2}, \frac{5\pi}{2}, \frac{9\pi}{2}, \dots$$

This is necessitated by such observations as the fact that

$$\pi/2 = \arg(i) = \arg[(-1)(-i)] = \arg(-1) + \arg(-i) = \pi + 3\pi/2 = 5\pi/2.$$

This fact, which appears to be a nuisance at this point will in fact turn out to be very useful in the next section. Some more light will be shed on this in Section 9.4. Theorem 2.2 is commonly referred to as the *argument principle*. It has many interesting and useful consequences. For example, it clearly implies that $\arg(z^2) = 2\arg(z)$ and $|z^2| = |z|^2$. Similarly, $\arg(z^3) = 3\arg(z)$ and $|z^3| = |z|^3$. In fact, a simple induction procedure yields the following observation.

Corollary 2.3 If z is any nonzero complex number and k is any positive integer, then $\arg(z^k) = k\arg(z)$ and $|z^k| = |z|^k$.

This observation can be put to good use in computing large powers of complex numbers. Consider the problem of computing $(1+i)^{100}$. By Corollary 2.3,

$$\arg[(1+i)^{100}] = 100\arg(1+i) = 100 \cdot (\pi/4) = 25\pi = \pi$$

and $|(1+i)^{100}| = |1+i|^{100} = (\sqrt{2})^{100} = 2^{50}$. Hence, $(1+i)^{100} = 2^{50}(\cos \pi + i \sin \pi) = -2^{50}$.

Corollary 2.3 also holds for nonpositive exponents if we define $z^0 = 1$ for all z and $z^{-k} = (1/z)^k$ for $z \neq 0$ and $k = 1, 2, 3, \dots$

The proof of this fact is relegated to Exercise 2.1.29. Exercise 2.1.28 calls for proving that integer powers of complex numbers obey the same rules as do the more familiar powers of real numbers, to wit, $z^m z^n = z^{m+n}$ and $(z^m)^n = z^{mn}$.

If $z = a + bi$ is any complex number, where a and b are real, we define \bar{z} , the *conjugate* of z , to be $a - bi$. Thus, $\overline{-2 + 3i} = -2 - 3i$ and $\overline{3 - 4i} = 3 + 4i$.

Theorem 2.4 If $z = a + bi$ is any complex number,

- (a) z and \bar{z} are symmetrical with respect to the x -axis;
- (b) $z\bar{z} = |z|^2$; $\arg(z\bar{z}) = \arg(z) + \arg(\bar{z})$;
- (c) $\overline{z + w} = \bar{z} + \bar{w}$; $\overline{zw} = \bar{z}\bar{w}$; $\bar{z}^{-1} = \overline{z^{-1}}$;
- (d) $\bar{z} = z$ if and only if z is real.

Proof. See Exercise 2.1.27. ■

Exercises 2.1

Find the argument and modulus of each of the complex numbers in Exercises 2.1.1 to 2.1.4

- 1. $2 + 3i$
- 2. $3 - 2i$
- 3. $-3 - 4i$
- 4. $-1 + 7i$

Express the complex quantities in Exercises 2.1.5 to 2.1.21 in the form $a + bi$, where a and b are real numbers.

- 5. $(2 + 3i) + (5 - i)$
- 6. $(17 - 3i) + (2 + 3i)$
- 7. $(2 + 3i)(5 - i)$
- 8. $(17 - 3i) - (2 + 3i)$
- 9. $(2 + 3i)(5 - i)$
- 10. $(17 - 3i)(2 + 3i)$
- 11. $(2 + 3i)/(5 - i)$
- 12. $(17 - 3i)/(2 + 3i)$
- 13. $(\sqrt{3} + 5i)/(2 - \sqrt{3}i)$
- 14. $(a + bi)/(a - bi) - (a - bi)/(a + bi)$
- 15. $(2 - i)^2/(1 + i)$
- 16. $(1 + i)^4$
- 17. $(1 - 2i)^4$
- 18. $(1 - i)^{63}$
- 19. $i^{4,321}$
- 20. $((1 - i)/(1 + i))^{127}$
- 21. i^{4n+3} (n is an integer)

Solve the equations in Exercises 2.1.22 to 2.1.25 for z and w :

- 22. $(1 + 2i)z + 5 = 0$
- 23. $(1 + i)z + 5i = \frac{z}{1-i} - 2$
- 24. $iz - w = 1 + i$ and $(1 + i)z + iw = 1$
- 25. $(1 - i)z + iw = i$ and $2z - (2 + i)w = 1$
- 26. Prove that if z and w are any two complex numbers, then $|z + w| \leq |z| + |w|$.

27. Prove the following:

(a) Theorem 2.4(a)

(c) Theorem 2.4(c)

(b) Theorem 2.4(b)

(d) Theorem 2.4(d)

(e) If a , b , c , and d are any real numbers, prove that z is a root of the equation $ax^3 + bx^2 + cx + d = 0$ if and only if \bar{z} is also a root of the same equation.

28. Prove that if z is any complex number and m and n are any integers, then $z^m z^n = z^{m+n}$ and $(z^m)^n = z^{mn}$.

29. Prove that Corollary 2.3 holds for every integer k .

30. Prove that the three distinct complex numbers z_1 , z_2 , and z_3 are collinear if and only if there exists a real number λ such that $z_2 = (1 - \lambda)z_1 + \lambda z_3$. (Hint: examine the expression $(z_2 - z_1)/(z_3 - z_1)$.)

31. Prove that if z and w are two complex numbers, then the distance between them equals $|z - w|$.

32. Prove that the midpoint of the line segment joining the complex numbers z and w is $(z + w)/2$.

33. Prove that the four complex numbers z_1 , z_2 , z_3 , and z_4 lie on either a common straight line or a common circle if and only if the number

$$\frac{(z_1 - z_3)/(z_1 - z_2)}{(z_4 - z_3)/(z_4 - z_2)}$$

is real.

34. Let z_1 , z_2 , z_3 , and z_4 be four complex numbers such that $|z_1| = |z_2| = |z_3| = |z_4| = 1$. Prove that z_1 , z_2 , z_3 , and z_4 form a rectangle if and only if $z_1 + z_2 + z_3 + z_4 = 0$.

35. Prove that the center of gravity of the triangle whose vertices are the complex numbers z_1 , z_2 , and z_3 is $(z_1 + z_2 + z_3)/3$. (Hint: recall that the center of gravity of a triangle coincides with the intersection of its three medians.)

36. Prove that if $|\xi| = 1$, then there is a real number b such that $(1 + \xi)/(1 - \xi) = bi$.

37. Prove that if $z = a + bi$, where a and b are real, then $(|a| + |b|)/\sqrt{2} \leq |z| \leq |a| + |b|$.

2.2 Complex Roots

In the previous section the four arithmetical operations were extended to complex numbers. Next we examine the process of finding roots of complex numbers. What, for example, is $\sqrt[4]{1}$? Before addressing this question, it behooves us to recall that even $\sqrt{1}$ involves some ambiguities. Sometimes it is 1 and sometimes it is -1 or both ± 1 . We therefore define $\sqrt[n]{z}$, for any complex number z and for any positive integer n , to be the set of all the complex numbers w such that $w^n = z$.

Returning to $\sqrt{1}$, let j be any complex number such that $j^2 = i$. Then $2\arg(j) = \arg(i)$ and $|j|^2 = |i| = 1$. Consequently, using $\arg(i) = \pi/2$, $\arg(j) = \arg(i)/2 = \pi/4$ and since $|j|$ is, by definition, positive, $|j| = 1$. Thus

$$j = 1 \left(\cos \frac{\pi}{4} + i \sin \frac{\pi}{4} \right) = \frac{\sqrt{2}}{2} + \frac{\sqrt{2}}{2}i.$$

Another square root of i is of course

$$-j = -\frac{\sqrt{2}}{2} - \frac{\sqrt{2}}{2}i.$$

An alternate method for arriving at $-j$ is to recall that $\arg(i)$ could also have been taken as $5\pi/2$, in which case we obtain the square root

$$1 \left(\cos \frac{5\pi}{4} + i \sin \frac{5\pi}{4} \right) = -\frac{\sqrt{2}}{2} - \frac{\sqrt{2}}{2}i = -j.$$

It can be easily verified by direct calculations that

$$\left(\frac{\sqrt{2}}{2} + \frac{\sqrt{2}}{2}i \right)^2 = i = \left(-\frac{\sqrt{2}}{2} - \frac{\sqrt{2}}{2}i \right)^2.$$

This procedure yields three different values for $\sqrt[3]{1}$. For, taking as the argument of 1 the successive values of $0, 2\pi, 4\pi, 6\pi, \dots$, each of the elements of $\sqrt[3]{1}$ must have as its argument one of the values $0, 2\pi/3, 4\pi/3, 2\pi, \dots$. The modulus of 1 being 1, it follows that the modulus of $\sqrt[3]{1}$ must be the real cube root of 1 which is also 1. Hence we get as

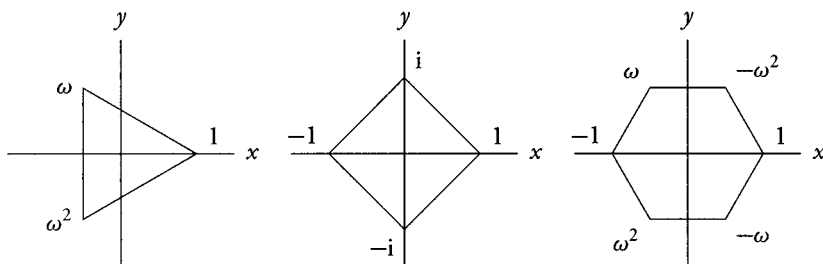


Figure 2.4 Complex roots of unity

our cube roots of 1 the following numbers:

$$1(\cos 0 + i \sin 0) = 1 \cdot 1 + i \cdot 0 = 1,$$

$$1\left(\cos \frac{2\pi}{3} + i \sin \frac{2\pi}{3}\right) = -\frac{1}{2} + \frac{\sqrt{3}}{2}i,$$

$$1\left(\cos \frac{4\pi}{3} + i \sin \frac{4\pi}{3}\right) = -\frac{1}{2} - \frac{\sqrt{3}}{2}i,$$

$$1(\cos 2\pi + i \sin 2\pi) = 1,$$

$$\vdots$$

It is clear that this list of cube roots of 1 will cycle through the same three values. The second root, the one with argument $2\pi/3$, is denoted by ω . By Corollary 2.3, the third root, having double the argument of ω and the same modulus of 1, equals ω^2 . Note that the Cartesian representations of these three complex cube roots of 1 form an equilateral triangle in the Cartesian plane (Figure 2.4).

The same procedure yields all the n -th roots of unity (roots of 1) for each positive integer n .

Theorem 2.5 Let n be a positive integer, and let $\zeta = \cos(2\pi/n) + i \sin(2\pi/n)$. Then $\sqrt[n]{1} = \{1, \zeta, \zeta^2, \zeta^3, \dots, \zeta^{n-1}\}$.

Proof. By Corollary 2.3, $\zeta^n = \cos 2\pi + i \sin 2\pi = 1$, and so ζ is indeed one of the elements of $\sqrt[n]{1}$. Moreover, for any integer k ,

$$(\zeta^k)^n = (\zeta^n)^k = 1^k = 1,$$

and hence each ζ^k is indeed an n -th root of unity. Since $\arg(\zeta^k) = 2\pi k/n$, it follows that the numbers $1, \zeta, \zeta^2, \dots, \zeta^{n-1}$ all have distinct arguments and so they are all distinct complex numbers. The remainder of the proof, that there are no other roots of unity, is relegated to Exercise 2.2.29. We note that this also follows from the fact (see Proposition 6.8) that a polynomial equation of degree n has at most n roots. ■

The n -th root of unity

$$\zeta = \cos \frac{2\pi}{n} + i \sin \frac{2\pi}{n}$$

will be referred to as the *first* n -th root of unity. It is clear that $|\zeta| = 1$, and hence, by Theorem 2.2, $|\zeta^k| = 1$ for every integer k . Since the angle subtended by the consecutive roots ζ^{k+1} and ζ^k at the origin equals

$$\arg\left(\frac{\zeta^{k+1}}{\zeta^k}\right) = \arg(\zeta) = \frac{2\pi}{n}$$

and is independent of k , it follows that the elements of $\sqrt[n]{1}$ form a regular n -gon that is centered at the origin. This fact is crucial for the next section, and so we state it as a proposition.

Proposition 2.6 For any fixed integer $n \geq 3$, the Cartesian representations of the elements of $\sqrt[n]{1}$ form the vertices of a regular n -gon.

For example, since $2\pi/4 = \pi/2$ and

$$\cos \frac{\pi}{2} + i \sin \frac{\pi}{2} = 0 + i = i,$$

it follows that

$$\sqrt[4]{1} = \{1, i, i^2, i^3\} = \{1, i, -1, -i\},$$

the elements of which form the vertices of a square (Figure 2.4).

Similarly, since $2\pi/6 = \pi/3$ and

$$\cos \frac{\pi}{3} + i \sin \frac{\pi}{3} = \frac{1}{2} + i \frac{\sqrt{3}}{2} = -\omega^2,$$

it follows that

$$\sqrt[6]{1} = \{1, -\omega^2, (-\omega^2)^2, (-\omega^2)^3, (-\omega^2)^4, (-\omega^2)^5\} = \{1, -\omega^2, \omega, -1, \omega^2, -\omega\},$$

the elements of which form the vertices of the regular hexagon of Figure 2.4.

The following proposition is both a natural extension and a corollary of Theorem 2.5. Since our main interest lies in the roots of unity, its proof is omitted and relegated to Exercise 2.2.18. The subsequent example clarifies the purport of the proposition.

Proposition 2.7 Let n be a positive integer and z any nonzero complex number with argument θ . If

$$\zeta = \cos \frac{2\pi}{n} + i \sin \frac{2\pi}{n},$$

then

$$\sqrt[n]{z} = \left\{ |\sqrt[n]{z}| \left(\cos \frac{\theta}{n} + i \sin \frac{\theta}{n} \right) \zeta^k \mid k = 0, 1, 2, \dots, n-1 \right\}$$

where $|\sqrt[n]{z}|$ denotes the common modulus of all the elements of $\sqrt[n]{z}$.

Since $-1 + i$ has modulus 2 and argument $3\pi/4$, and since

$$\cos \frac{\pi}{4} + i \sin \frac{\pi}{4} = \frac{1}{\sqrt{2}}(1 + i),$$

it follows that

$$\begin{aligned} \sqrt[3]{-1+i} &= \left\{ \sqrt[3]{\sqrt{2}} \left(\frac{1+i}{2} \right), \sqrt[3]{\sqrt{2}} \left(\frac{1+i}{2} \right) \omega, \sqrt[3]{\sqrt{2}} \left(\frac{1+i}{2} \right) \omega^2 \right\} \\ &= \left\{ \frac{1+i}{\sqrt[3]{2}}, \frac{1+i}{\sqrt[3]{2}} \omega, \frac{1+i}{\sqrt[3]{2}} \omega^2 \right\}. \end{aligned}$$

It should be noted here that calculators are very handy in computing complex roots too. Thus, to compute $\sqrt[4]{2+3i}$ we note that

$$\sqrt[4]{|2+3i|} = \sqrt[4]{\sqrt{2^2+3^2}} = \sqrt[8]{13} \approx 1.378$$

and

$$\arg(\sqrt[4]{2+3i}) = \frac{1}{4} \arg(2+3i) = \frac{1}{4} \arctan\left(\frac{3}{2}\right) \approx 14.077^\circ.$$

Hence, if we set

$$w = \cos \left[14 \arctan\left(\frac{3}{2}\right) \right] + i \sin \left[\frac{1}{4} \arctan\left(\frac{3}{2}\right) \right] \approx .970 + .243i,$$

then

$$\begin{aligned}\sqrt[4]{2+3i} &\approx 1.378 \{ w, wi, -w, -wi \} \\ &\approx 1.378 \{ .970 + .243i, -.243 + .970i, -.970 - .243i, .243 - .970i \} \\ &\approx \{ 1.336 + .335i, -.335 + 1.336i, -1.336 - .335i, .335 - 1.336i \}.\end{aligned}$$

The solution of the quadratic equation detailed in Chapter 1 works for complex coefficients as well. Accordingly, the roots of the equation $iz^2 + 2z - 2i = 0$ are

$$\frac{-2 \pm \sqrt{2^2 - 4 \cdot i \cdot (-2i)}}{2i} = \frac{-2 \pm \sqrt{-4}}{2i} = \frac{-2 \pm 2i}{2i} = i \pm 1.$$

We conclude this section with a curious fact that will shortly prove unexpectedly useful.

Proposition 2.8 For any fixed integer $n > 1$, the sum of the elements of $\sqrt[n]{1}$ is 0.

Proof. Let ζ be the first n -th root of unity. By Theorem 2.5, the elements of $\sqrt[n]{1}$ can be listed as $1, \zeta, \zeta^2, \dots, \zeta^{n-1}$. The formula for geometric progressions now yields

$$1 + \zeta + \zeta^2 + \zeta^3 + \dots + \zeta^{n-1} = \frac{1 - \zeta^n}{1 - \zeta} = \frac{1 - 1}{1 - \zeta} = 0. \quad \blacksquare$$

Exercises 2.2

Express each of the elements of the sets in Exercises 2.2.1 to 2.2.12 in the form $a + bi$, where a and b are real numbers.

1. $\sqrt[8]{1}$

7. $\sqrt[3]{1+i}$

2. $\sqrt[12]{1}$

8. $\sqrt[3]{i}$

3. $\sqrt[6]{-1}$

9. $\sqrt[3]{-i}$

4. $\sqrt[4]{-i}$

10. $\sqrt[5]{100-37i}$

5. $\sqrt{3-4i}$

11. $\sqrt{c^2-1+2ci}$

6. $\sqrt{8-30i}$

12. $\sqrt{4cd-2(c^2-d^2)i}$

13. Resolve the following paradox:

$$1 = \sqrt{1} = \sqrt{(-1)(-1)} = \sqrt{-1}\sqrt{-1} = i \cdot i = i^2 = -1.$$

Find the complex solutions of the equations in Exercises 2.2.14 to 2.2.17.

14. $z^2 - 6z + 9 + 2i = 0$

16. $z^2 - (1 + i)z + 5i = 0$

15. $(1 - 2i)z^2 + 2z + 1 = 0$

17. $z^2 + 3(1 + i)z - (2 - 3i) = 0$

18. Prove Proposition 2.7.

Prove the identities in Exercises 2.2.19 to 2.2.22.

19. $(1 + \omega^2)^{16} = \omega$

20. $(3 + 5\omega + 3\omega^2)^9 = 512$

21. $a^3 + b^3 = (a + b)(a + b\omega)(a + b\omega^2)$

22. $(a + b + c)(a + b\omega + c\omega^2)(a + b\omega^2 + c\omega) = a^3 + b^3 + c^3 - 3abc$

If ζ is an n -th root of unity, simplify ζ^k for the values of n and k specified in Exercises 2.2.23 to 2.2.26.

23. $n = 10, k = 135$

25. $n = 999, k = 12,345$

24. $n = 135, k = 999$

26. $n = 12,345, k = 106$

27. Prove that $z \in \sqrt[n]{1}$ if and only if $\bar{z} \in \sqrt[n]{1}$.

28. Prove that for every positive integer $n > 1$,

$$\sum_{k=1}^n \cos \frac{2\pi k}{n} = 0 = \sum_{k=1}^n \sin \frac{2\pi k}{n}.$$

29. Complete the proof of Theorem 2.5 by showing that the given list contains all the n -th roots of 1.

30. Prove that three complex numbers A , B , and C form a counterclockwise equilateral triangle if and only if $A + B\omega + C\omega^2 = 0$. What equation characterizes clockwise equilateral triangles?

31. Suppose ABC is any triangle in the plane. Let A' , B' , and C' be three points such that $A'BC$, $AB'C$, and ABC' are all clockwise (or all counterclockwise) equilateral triangles. Prove that the centers of triangles $A'BC$, $AB'C$, and ABC' also form an equilateral triangle.

2.3 Solvability by Radicals I

We now have sufficient tools at our disposal to formalize the notion of an *algebraic solution* of an equation

$$a_0 x^n + a_1 x^{n-1} + \cdots + a_{n-1} x + a_n = 0$$

where a_0, a_1, \dots, a_n are any complex numbers. A solution of this equation is of course another complex number r such that $a_0 r^n + a_1 r^{n-1} + \cdots + a_{n-1} r + a_n = 0$. The value of the solution r clearly depends on the coefficients a_0, a_1, \dots, a_n , and the solution is said to be algebraic if this dependence involves only radicals and the four arithmetic operations. More precisely, let \mathbb{Z} denote the set of integers and let V be any set of complex numbers. The complex number z is said to have an *algebraic expression* in V if there exists a sequence of complex numbers $z_1, z_2, \dots, z_n = z$ such that for each $i = 1, 2, \dots, n$ either $z_i \in \sqrt[m]{z_{i-1}}$ for some positive integer $m > 1$ or else the number z_i is obtained by adding, subtracting, multiplying, or dividing some elements of $\mathbb{Z} \cup V \cup \{z_1, z_2, \dots, z_{i-1}\}$. Thus, each of the solutions of the quadratic equation $ax^2 + bx + c = 0$ (where $a \neq 0$), namely

$$z = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

has an algebraic expression in $\{a, b, c\}$ because it is possible to choose

$$\begin{array}{llll} z_1 = b^2, & z_2 = ac, & z_3 = 4z_2, & z_4 = z_1 - z_3, \\ z_5 \in \sqrt{z_4}, & z_6 = -b + z_5 & z_7 = z_6/a, & z = z_8 = z_7/2. \end{array}$$

If the complex number z has an algebraic expression in V whose only radicals are square roots, then we say that z has a degree 2 algebraic expression in V . Thus, the above solutions of the quadratic equation clearly have degree 2 radical expressions in $\{a, b, c\}$. If no radicals appear in an algebraic expression of z in V , then z is said to have a rational expression in V . For example, if $c \neq -1$, then $(2 - ab)/(1 + c)$ has a rational expression in $\{a, b, c\}$ with $n = 4$ where $z_1 = ab$, $z_2 = 2 - z_1$, $z_3 = 1 + c$, and $z = z_4 = z_2/z_1$. If z has an algebraic expression in V , and V happens to be empty, we shall say that z has an algebraic expression in the integers. Finally, we say that an equation

$$a_0 x^n + a_1 x^{n-1} + a_2 x^{n-2} + \cdots + a_n = 0$$

is *solvable by radicals* or *algebraically resolvable* if each of its roots has an algebraic expression in the coefficients $\{a_0, a_1, a_2, \dots, a_n\}$. Thus, the quadratic equation $ax^2 + bx + c = 0$ is solvable by radicals, as is easily verified by examining the quadratic formula above. The equation

$$x^4 + ax^3 + bx^2 + ax + 1 = 0 \quad (2.9)$$

is also solvable by radicals. To see this we observe that 0 is not a root of this equation so that it can be divided by x^2 and its terms can be regrouped as

$$x^2 + \frac{1}{x^2} + a\left(x + \frac{1}{x}\right) + b = 0$$

or

$$\left(x + \frac{1}{x}\right)^2 + a\left(x + \frac{1}{x}\right) + b - 2 = 0.$$

Setting $u = x + 1/x$ we note that $u^2 + au + b - 2 = 0$ and $x^2 - ux + 1 = 0$. Thus, x has an algebraic expression in $\{u\}$, and u , in turn, has an algebraic expression in $\{a, b\}$. This, of course, means that every solution of Equation 2.9 has an algebraic expression in its coefficients.

The following observations are easily proved by induction on the minimum number of equations needed to express a number v as an algebraic expression of a set W :

- If x has a rational expression in V and each element v of V has a rational expression in the set W , then v has a rational expression in the elements of W .
- If x has an algebraic expression in V and each element v of V has an algebraic expression in the set W , then v has an algebraic expression in the elements of W .
- If x has an algebraic expression in V and each element v of V has a rational expression in the set W , then v has an algebraic expression in the elements of W .
- If x has a rational expression in V and each element v of V has an algebraic expression in the set W , then v has an algebraic expression in the elements of W .

Exercises 2.3

Decide whether the expressions in Exercises 2.3.I to 2.3.II are rational or (degree 2) algebraic expressions in $\{x, y, z\}$.

- | | |
|-------------------------------------|-------------------------------------|
| 1. $(x + y - 1)/(x - z + 2)$ | 3. $(x + y - 1)/(x - z + \sqrt{2})$ |
| 2. $(x + \sqrt{y} - 1)/(x - z + 2)$ | 4. $(x + y)^{10^{10}}$ |

22. Prove that the roots of the equation $(A + C - B)x^2 + 2Cx + (B + C - A) = 0$ have rational expressions in $\{A, B, C\}$.
23. Prove that the roots of the equation $ABC^2x^2 + (3A^2 + B^2)Cx - (6A^2 + AB - 2B^2) = 0$ have rational expressions in $\{A, B, C\}$.

2.4 Ruler-and-Compass Constructibility of Regular Polygons

The ancient Greek mathematicians, who invented what we have come to call Euclidean geometry and the notion of a rigorous proof, bequeathed their successors a host of unsolved mathematical problems. Best-known amongst these are the questions of whether it is possible to trisect an angle, double a cube, or square a circle by means of a compass and an unmarked ruler alone. Here we treat a lesser-known, but equally natural, construction problem, namely, what regular polygons are constructible by ruler and compass alone? The other three problems are discussed informally at the end of the section.

The ruler-and-compass constructions of the equilateral triangle, the square, and the regular hexagon are standard fare in the high school curriculum. That the regular pentagon is also so constructible is true, but not so widely known. This is proved below in Proposition 2.14. A regular octagon is easily constructed by inscribing a square in a circle and then drawing the two diameters that are perpendicular to the sides of the square (Figure 2.5). In general, it is clear that given any regular n -gon it is possible to derive from it a regular $2n$ -gon by drawing radii perpendicular to its sides. Hence the regular n -gon is constructible for $n = 2^{m+2}$, $3 \cdot 2^m$, and $5 \cdot 2^m$ for $m = 0, 1, 2, \dots$. If a regular pentagon and an equilateral triangle are inscribed in a circle so that they share a vertex, as in Figure 2.6, then arc AB is $2/5 - 1/3 = 1/15$ of the total circumference of the circle. It follows that the regular 15-sided polygon is also constructible by ruler and compass. This information is summarized as the following proposition.

Proposition 2.10 The regular n -sided polygon is constructible by ruler and compass for $n = 3, 4, 5, 6, 8, 10, 12, 15, 16$.

In 1796, Gauss proved that the regular 17-sided polygon is also constructible by ruler and compass alone, and we shall discuss in detail a proof that appears in his *Disquisitiones Arithmeticae* of 1801. Before that, however, it is necessary to make some general remarks about ruler-and-compass constructibility. We shall work in the Cartesian plane, so it will be assumed that a certain line segment has been designated as the *unit segment*. For the sake of brevity it will henceforth be said that a configuration is *constructible* when it is constructible by ruler and compass alone. A real number α is said to be *constructible* if it

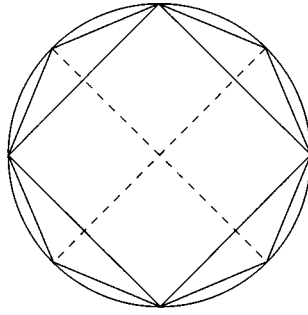


Figure 2.5 A square and a regular octagon

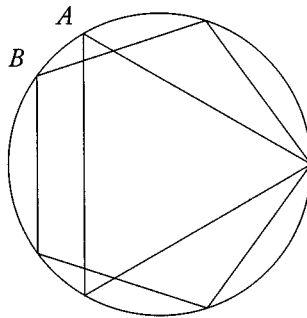


Figure 2.6 An equilateral triangle and a regular pentagon

is possible to construct a line segment whose length is $|\alpha|$ times the length of the unit segment.

Proposition 2.11 The point (x, y) is constructible if and only if its coordinates x and y are constructible real numbers.

Proof. The nature of Cartesian coordinates is such that it is possible to pass from a point to its coordinates and vice versa by means of straight lines that are perpendicular to the axes. Since such perpendiculars are well known to be constructible, we are done. ■

It is obvious that the number 1 is constructible and it follows from elementary Euclidean geometry that if α and β are constructible real numbers, so are $\alpha \pm \beta$. In particular, every integer is constructible. The next lemma will provide us with a host of real constructible numbers.

Lemma 2.12 If α and β are nonzero constructible real numbers, then so are their product $\alpha\beta$, their quotient β/α , and the square root $\sqrt{|\alpha|}$.

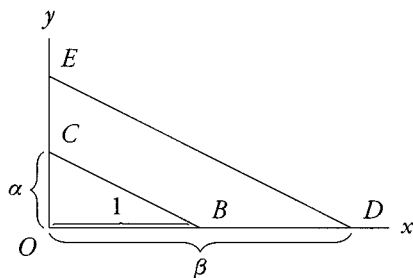


Figure 2.7 The multiplication of constructible numbers

Proof. It is clear that we may restrict attention to positive α and β . On the positive x - and y -axes (Figure 2.7) let B , C , and D be points such that the lengths of the segments OB , OC , and OD are 1, α , and β , respectively. Using a standard ruler-and-compass construction, draw through D a straight line that is parallel to BC and intersects OY at E . Since $\triangle OBC$ and $\triangle ODE$ are similar, it follows that $OD/OB = OE/OC$, or $\beta/1 = OE/\alpha$, and so the constructible line segment OE has length $\alpha\beta$.

In view of the above argument it suffices to show that $1/\alpha$ is constructible. On the positive x - and y -axes (Figure 2.8) let B , D , and E be points such that the lengths of the segments OB , OD , and OE are 1, α , and 1, respectively. Join the points D and E and draw a line through B that is parallel to DE . If this line intersects OY at the point C , then, because of the similarity of $\triangle OBC$ and $\triangle ODE$, it follows that $OD/OB = OE/OC$ or $\alpha/1 = 1/OC$, and so the constructible line segment OC has length $1/\alpha$.

Let A , D , and B be three collinear points such that the line segments AD and DB have lengths α and 1 respectively (Figure 2.9). Let C be the intersection of the line perpendicular to AB at the point D with the semicircle that has AB as its diameter (all these are well known to be constructible). The triangles ACD and CBD are right triangles each of which also shares an acute angle with the right triangle $\triangle ABC$. Thus, $\triangle ACD$ and $\triangle CBD$ are both similar to $\triangle ABC$ and hence they are also similar to each other. Consequently, $AD/CD = CD/BD$, or $\alpha/CD = CD/1$, and hence the constructible segment CD has length a . ■

It follows from Lemma 2.12 that every rational number is constructible, as is every real number that has a degree 2 algebraic expression in the integers. A complex number is said to be constructible if its Cartesian representation is a constructible point. The above

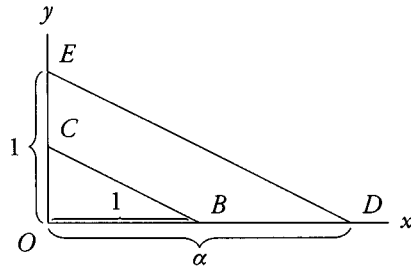


Figure 2.8 The reciprocal of a constructible number

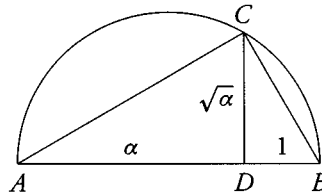


Figure 2.9 The square root of a constructible number

geometric proposition results in an algebraic description of some constructible complex numbers.

Corollary 2.13 If the complex number $z = x + iy$ has a degree 2 algebraic expression in the integers, then it is constructible.

Proof. Because of the recursive nature of the definition of algebraic expressions it suffices to show that if $z = x + iy$ and $w = u + iv$ are constructible complex numbers, then so are $z + w$, $z - w$, zw , z/w , and \sqrt{z} . However, if z and w are constructible complex numbers, then, by Proposition 2.11 and Lemma 2.12, so are $z \pm w = (x \pm u) + i(y \pm v)$, $zw = (xu - yv) + i(xv + yu)$,

$$\frac{1}{z} = \frac{x}{x^2 + y^2} - i \frac{y}{x^2 + y^2},$$

and $z/w = z(1/w)$.

To argue the constructibility of \sqrt{z} we note that it consists of the intersections of the circle centered at the origin and of radius $|z|$ with the straight line that bisects the argument of z . Since $|z|$ is constructible by Lemma 2.12, and angle bisection is a well-known ruler-and-compass construction, we are done. ■

The converse of this corollary also holds, and its proof is relegated to Exercise 2.4.28.

Our approach to proving the constructibility of the regular pentagon and the regular 17-sided polygon is based on the observation that the elements of $\sqrt[n]{1}$ form a regular n -gon centered at the origin of the Cartesian plane. By Corollary 2.13 it suffices to show that each of the vertices of these polygons, when regarded as a complex number, has a degree 2 algebraic expression in the integers.

Proposition 2.14 The regular pentagon is constructible by ruler and compass alone.

Proof. Let ε denote the first 5-th root of unity. Since the remaining roots are $\varepsilon^2, \varepsilon^3, \varepsilon^4$, and 1, it follows from Lemma 2.12 that it suffices to prove that ε has a degree 2 algebraic expression in the integers. By Proposition 2.8, ε satisfies the equation

$$\varepsilon + \varepsilon^2 + \varepsilon^3 + \varepsilon^4 = -1.$$

Set $A = \varepsilon + \varepsilon^4$ and $B = \varepsilon^2 + \varepsilon^3$, and note that

$$A + B = \varepsilon + \varepsilon^4 + \varepsilon^2 + \varepsilon^3 = -1$$

and

$$AB = (\varepsilon + \varepsilon^4)(\varepsilon^2 + \varepsilon^3) = \varepsilon^3 + \varepsilon^4 + \varepsilon^6 + \varepsilon^7 = \varepsilon^3 + \varepsilon^4 + \varepsilon + \varepsilon^2 = -1.$$

Hence, by Proposition 1.3, A and B are the solutions of the quadratic equation $x^2 + x - 1 = 0$, and so A has a degree 2 algebraic expression in the integers. On the other hand,

$$A = \varepsilon + \varepsilon^4 = \varepsilon + \frac{1}{\varepsilon},$$

so that $\varepsilon^2 - A\varepsilon + 1 = 0$ and hence ε is a solution of the quadratic equation $x^2 - Ax + 1 = 0$. It follows that ε has a degree 2 algebraic expression in A , which in turn has a degree 2 algebraic expression in the integers. Hence, by Corollary 2.13, ε is constructible. ■

Exercise 2.4.1 calls for the explicit description of the coordinates of ε as degree 2 algebraic expressions in the integers. Exercise 2.4.29 calls for a ruler-and-compass construction of the regular pentagon. We now turn to the regular 17-sided polygon (Figure 2.10).

Theorem 2.15 The regular 17-sided polygon is ruler-and-compass constructible.

Proof. Let $\zeta = \cos(2\pi/17) + i\sin(2\pi/17)$ be the first 17-th root of unity. Just as was the case for the regular pentagon, it suffices to show that ζ has a degree 2 algebraic expression

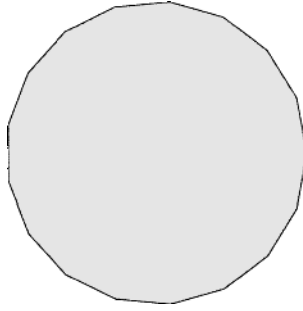


Figure 2.10 The regular 17-gon

in the integers. This will be accomplished by listing a sequence of elements, the last being ζ , such that each member of the sequence has a degree 2 algebraic expression in the previous ones, and the first has a degree 2 algebraic expression in the integers.

We begin with the not-at-all-natural observation that by Proposition 2.8,

$$\zeta + \zeta^3 + \zeta^9 + \zeta^{10} + \zeta^{13} + \zeta^5 + \zeta^{15} + \zeta^{11} + \zeta^{16} + \zeta^{14} + \zeta^8 + \zeta^7 + \zeta^4 + \zeta^{12} + \zeta^2 + \zeta^6 = -1.$$

Next, set

$$\begin{aligned} A &= \zeta + \zeta^9 + \zeta^{13} + \zeta^{15} + \zeta^{16} + \zeta^8 + \zeta^4 + \zeta^2, \\ B &= \zeta^3 + \zeta^{10} + \zeta^5 + \zeta^{11} + \zeta^{14} + \zeta^7 + \zeta^{12} + \zeta^6, \\ C &= \zeta + \zeta^{13} + \zeta^{16} + \zeta^4, \\ D &= \zeta^9 + \zeta^{15} + \zeta^8 + \zeta^2, \\ E &= \zeta^3 + \zeta^5 + \zeta^{14} + \zeta^{12}, \\ F &= \zeta^{10} + \zeta^{11} + \zeta^7 + \zeta^6, \\ G &= \zeta + \zeta^{16} = \zeta + \zeta^{-1}, \\ H &= \zeta^{13} + \zeta^4. \end{aligned}$$

It is not at all clear at this point what the pattern is for writing out the imaginary 17-th roots of unity in the first of the above equations, but once that is taken for granted, the pattern for forming the subsequent sequences is quite obvious.

Since $G = \zeta + \zeta^{-1}$, it follows that $\zeta^2 - G\zeta + 1 = 0$, and hence ζ has a degree 2 algebraic expression in G . Next, $G + H = \zeta + \zeta^{16} + \zeta^{13} + \zeta^4 = C$ and

$$GH = (\zeta + \zeta^{16})(\zeta^{13} + \zeta^4) = \zeta^{14} + \zeta^5 + \zeta^{29} + \zeta^{20} = \zeta^{14} + \zeta^5 + \zeta^{12} + \zeta^3 = E.$$

Hence, G and H are the roots of $x^2 - Cx + E = 0$.

Again, by making use of Proposition 2.8 it is easily verified (Exercise 2.4.3) that $C + D = A$, $CD = -1$, $E + F = B$, and $EF = -1$, from which it follows that C and D are the roots of $x^2 - Ax - 1 = 0$ and E and F are the roots of $x^2 - Bx - 1 = 0$. Finally, leaving some details to Exercise 2.4.4, we see that $A + B = -1$ and $AB = -4$ so that A and B are the roots of $x^2 + x - 4 = 0$. Thus,

- A and B have degree 2 algebraic expressions in the integers,
- C and E have degree 2 algebraic expressions in A and B ,
- G has a degree 2 algebraic expression in C and E , and
- ζ has a degree 2 algebraic expression in G .

Consequently, ζ has a degree 2 algebraic expression in the integers. Thus, the Cartesian representations of ζ and all of its powers are constructible. ■

While the above proof demonstrates the ruler-and-compass constructibility of the regular 17-sided polygon, it sidesteps the issue of actually constructing the vertices. The fact that the abscissa of the point ζ that is used in the proof of the above theorem is

$$-\frac{1}{16} + \frac{1}{16}\sqrt{17} + \frac{1}{16}\sqrt{34 - 2\sqrt{17}} + \frac{1}{8}\sqrt{17 + 3\sqrt{17} - \sqrt{34 - 2\sqrt{17}} - 2\sqrt{34 + 2\sqrt{17}}}$$

clarifies the reasons behind this avoidance.

Gauss did not limit himself to the 17-gons. He went on to prove that for each prime number p the equation $x^p - 1 = 0$ has a resolution of the type exhibited in Proposition 2.14 and Theorem 2.15. More specifically, Gauss demonstrated that the roots of $x^p - 1 = 0$ have algebraic expressions in the integers that call only for radicals of the type $\sqrt[q]{x}$ where q is any prime factor of $p - 1$. In particular, if $p - 1$ is a power of 2, as is the case for $p = 5$ and $p = 17$, then these roots all have degree 2 algebraic expressions in the integers, as illustrated by Proposition 2.14 and Theorem 2.15.

This, of course, means that for such p the regular p -sided polygon is constructible by ruler and compass. It so happens (Exercise 2.4.5) that if p is a prime such that $p - 1$ is a

power of 2, then $p = 2^{2^m} + 1$ for some integer m . Unfortunately, only five such values of p are known, namely

$$3 = 2^{2^0} + 1, \quad 5 = 2^{2^1} + 1, \quad 17 = 2^{2^2} + 1, \quad 257 = 2^{2^3} + 1, \quad 65,537 = 2^{2^4} + 1.$$

The next number in this sequence,

$$2^{2^5} + 1 = 4,294,967,297$$

is not a prime since it equals $641 \cdot (6,700,417)$, as discovered by Euler in the eighteenth century. In fact, all of the numbers

$$2^{2^m} + 1$$

for $m = 5, 6, \dots, 16$ are now known to be composite.

Gauss also claimed to have proved rigorously that when p is a prime that is not of the form $2^n + 1$ (such as 7 or 11), the regular p -sided polygon is not constructible by ruler and compass alone. While he never wrote down his proof, this fact is now known to be true.

The strange ordering of the 17-th roots of unity that was used in the proof of Theorem 2.15 merits some discussion. The series

$$\zeta + \zeta^3 + \zeta^9 + \zeta^{10} + \zeta^{13} + \zeta^5 + \zeta^{15} + \zeta^{11} + \zeta^{16} + \zeta^{14} + \zeta^8 + \zeta^7 + \zeta^4 + \zeta^{12} + \zeta^2 + \zeta^6$$

is such that each root is the cube of the previous one. Thus,

$$(\zeta^3)^3 = \zeta^9, \quad (\zeta^9)^3 = \zeta^{27} = \zeta^{10}, \quad \dots, \quad (\zeta^2)^3 = \zeta^6.$$

The general implications of the existence of such a sequence, those which Gauss used to treat $x^p - 1 = 0$, lie beyond the scope of this text. However, it is necessary to point out that the existence of such a sequence cannot be taken for granted. For example, if instead of cubing each term we only squared each term, we would obtain the series

$$\zeta + \zeta^2 + \zeta^4 + \zeta^8 + \zeta^{16} + \zeta^{15} + \zeta^{13} + \zeta^9 + \zeta + \dots,$$

which fails to contain all the imaginary 17-th roots of unity. That some such power does cycle through all the imaginary p -th roots of unity will eventually be established in Theorem 7.17. At this point we leave the determination of such powers to trial and error. For example, if $p = 11$, then squaring works, since

$$\zeta + \zeta^2 + \zeta^4 + \zeta^8 + \zeta^5 + \zeta^{10} + \zeta^9 + \zeta^7 + \zeta^3 + \zeta^6$$

contains all the imaginary 11-th roots of unity whereas cubing does not work as the series

$$\zeta + \zeta^3 + \zeta^9 + \zeta^5 + \zeta^4 + \zeta + \dots$$

comes up short.

Three other constructibility problems were mentioned in this section's opening paragraph. Using tools similar to those that were introduced here, together with some linear algebra, it can be shown that the ruler-and-compass constructions in question do not exist. Specifically, the squaring of a circle calls for the ruler-and-compass construction of a square whose area equals that of a given circle. In particular, squaring a circle of radius 1 is tantamount to solving the equation $\pi \cdot 1^2 = x^2$ and so, if successful, such a construction would imply that $x = \sqrt{\pi}$ has a degree 2 expression in the integers. However, in 1882, Ferdinand Lindemann (1852–1939) proved that π has no algebraic expressions in the integers whatsoever. Consequently, neither does $\sqrt{\pi}$ have such an expression and hence there is no general ruler-and-compass construction for squaring arbitrary circles. Numbers like π that have no algebraic expressions in the integers are said to be *transcendental numbers*. Other transcendental numbers are $e = 2.718281\dots$ and

$$0.1234567891011121314151617181920212223\dots$$

The *doubling of a cube* calls for the ruler-and-compass construction of the side of a cube that has double the volume of a given cube. In particular the doubling of the unit cube, if successful, would prove that $\sqrt[3]{2}$ is a constructible number. Again, it is known that $\sqrt[3]{2}$ has no degree 2 algebraic expressions in the integers, and so this construction too is impossible.

Finally, the *trisection of an angle* calls for the ruler-and-compass constructions of an angle that is one-third of a given angle. In particular, if successful, such a construction would yield an angle of $20^\circ = 60^\circ/3$. This in turn would imply that $\cos 20^\circ$ is a constructible

number. However,

$$1/2 = \cos 60^\circ = \cos(3 \cdot 20^\circ) = 4 \cos^3 20^\circ - 3 \cos 20^\circ,$$

and so $\cos 20^\circ$ is a solution of the cubic equation $4x^3 - 3x - 1/2 = 0$. The solutions of this equation can be shown to have no degree 2 algebraic expressions in the integers. Consequently, there is no general ruler-and-compass construction for trisecting angles.

The credit for hammering the final nail into the coffins of the angle trisection and cube doubling problems in 1837 goes to the little-known French mathematician Pierre Wantzel (1814–1848).

Exercises 2.4

1. Show that if $\varepsilon = \cos 2\pi/5 + i \sin 2\pi/5$, then

$$\varepsilon = \frac{\sqrt{5}-1}{4} + \frac{\sqrt{10+2\sqrt{5}}}{4}i.$$

2. With ε as in Exercise 2.4.1, express ε^2 , ε^3 , and ε^4 in the form $a + bi$.
3. If C , D , E , and F are as defined in the proof of Theorem 2.15, prove that $CD = EF = -1$.
4. If A and B are as defined in the proof of Theorem 2.15, prove that $AB = -4$.
5. Prove that if p is a prime integer such that $p - 1$ is a power of 2, then there exists an integer m such that $p = 2^{2^m} + 1$.
6. For which of the integers $n = 3, 4, \dots, 100$ can you assert that a regular n -gon is constructible?
7. For which of the integers $n = 101, 102, \dots, 200$ can you assert that a regular n -gon is constructible?

Sketch diagrams that demonstrate the constructibility of the real numbers in Exercises 2.4.8 to 2.4.11.

8. $\sqrt{2}$

10. $\sqrt{\sqrt{2} + 5}$

9. $\sqrt{5}$

11. $\sqrt[4]{2}$

Suppose you are given a triangle with constructible sides. Which of the following aspects of this triangle, listed in Exercises 2.4.12 to 2.4.20, can you assert to be constructible? Justify your answers.

12. the perimeter
13. the square of the perimeter
14. the sum of the lengths of the medians
15. the square root of the sum of the lengths of the medians
16. the area
18. the cube of the area
17. the square of the area
19. the cube root of the area
20. the product of its area with its perimeter
21. Prove that the area of the regular hexagon whose side is constructible is a constructible real number.
22. Prove that the area of the regular pentagon whose side is constructible is a constructible real number.
23. Explain why the equation $x^8 + x^4 + 1 = 0$ has no real roots and why each of its complex roots is constructible.
24. Explain why the equation $x^{1,024} + x^{512} + 1 = 0$ has no real roots and why each of its complex roots is constructible.
25. Let η be the first 7-th root of unity. Prove that the quantities $\eta + \eta^2 + \eta^4$ and $\eta^3 + \eta^6 + \eta^5$ are constructible.
26. Let α be the first 11-th root of unity. Prove that the sum of α , α^3 , α^4 , α^5 , and α^9 is a constructible complex number.
27. Find three distinct 13-th roots of unity whose sum is a constructible number.
28. State and prove the converse of Corollary 2.13.
29. Construct a regular pentagon by means of ruler and compass.

2.5 Orders of Roots of Unity

We have seen that the 4-th roots of unity are 1, i , -1 , and $-i$ and that the 6-th roots of unity are 1, $-\omega^2$, ω , -1 , ω^2 , and $-\omega$. However, -1 is already a square root of 1, and ω and ω^2 are also cube roots of 1. If ζ is any root of unity, then the *order* of ζ , denoted by $o(\zeta)$, is the least positive integer m such that $\zeta^m = 1$. Thus, $o(1) = 1$, $o(-1) = 2$, $o(\omega) = 3$, $o(-\omega) = 6$, $o(i) = 4$, and $o(-\omega^2) = 6$.

The following proposition on the order of roots may seem obvious, but it does require formal proof. The integer m is said to be a *divisor* of the integer n (and n is said to be a *multiple* of m) if there is an integer k such that $n = km$, denoted by $m \mid n$. An integer that is greater than 1 and whose only positive divisors are 1 and itself is said to be *prime*. An integer that is greater than 1 and is not a prime is said to be *composite*.

Proposition 2.16 If ζ is any complex root of unity and n is any integer, then $\zeta^n = 1$ if and only if n is a multiple of $o(\zeta)$.

Proof. If n is a multiple of $o(\zeta)$, then there exists an integer m such that $n = o(\zeta)m$ and hence $\zeta^n = (\zeta^{o(\zeta)})^m = 1^m = 1$. Conversely, suppose that n is an integer such that $\zeta^n = 1$. If n is positive then the process of long division yields integers q and r such that $q \geq 0$, $o(\zeta) > r \geq 0$, and $n = o(\zeta)q + r$. But then

$$\zeta^r = \zeta^{n-o(\zeta)q} = \frac{\zeta^n}{(\zeta^{o(\zeta)})^q} = \frac{1}{1^q} = 1.$$

Since $0 \leq r < o(\zeta)$ and $o(\zeta)$ is the least positive integer m such that $\zeta^m = 1$, it follows that $r = 0$ and hence $n = o(\zeta)q$.

If n is zero, then it is trivially a multiple of $o(\zeta)$. If n is negative, then $-n$ is a positive integer such that

$$\zeta^{-n} = \frac{1}{\zeta^n} = 1.$$

Thus, by the above considerations, $o(\zeta)$ is a divisor of $-n$ and therefore also of n . ■

The following corollary is an immediate consequence of the above proposition. Its proof is relegated to Exercise 2.5.17.

Corollary 2.17 Suppose ζ is a root of unity and a and b are any two integers. Then

- $\zeta^a = \zeta^b$ if and only if $o(\zeta)$ is a divisor of $a - b$, and
- $1, \zeta, \zeta^2, \dots, \zeta^{o(\zeta)-1}$ are all distinct.

A primitive n -th root of unity is one which is not an m -th root for any $m < n$. Thus, i and $-i$ are the only primitive 4-th roots of unity, and $-\omega$ and $-\omega^2$ are the only primitive 6-th roots of unity. On the other hand, ω and ω^2 are primitive cube roots of unity, and, similarly, every 5-th root of unity except 1 is also a primitive fifth root of unity. It is clear that ζ is a primitive n -th root of unity if and only if $n = o(\zeta)$. It also follows from Corollary 2.17 that ζ is a primitive n -th root of unity if and only if $\zeta^n = 1$ and the numbers $1, \zeta, \zeta^2, \dots, \zeta^{n-1}$ are all distinct. In particular, the first n -th root of unity is always primitive.

We shall state and prove several more important facts regarding the orders of roots of unity in Section 5.2, after some necessary tools have been acquired.

Exercises 2.5

For each of the values of n in Exercises 2.5.1 to 2.5.8 list the elements of $\sqrt[n]{1}$ with their orders.

- | | | | |
|------|------|------|-------|
| 1. 4 | 3. 6 | 5. 8 | 7. 10 |
| 2. 5 | 4. 7 | 6. 9 | 8. 12 |

For each of the values of n in Exercises 2.5.9 to 2.5.12 list the primitive elements of $\sqrt[n]{1}$.

- | | | | |
|------|--------|--------|--------|
| 9. 6 | 10. 11 | 11. 12 | 12. 24 |
|------|--------|--------|--------|

13. Prove that if ζ is a root of unity, then $\text{o}(\zeta^{-1}) = \text{o}(\zeta)$.
14. Prove that if ζ is a primitive n -th root of unity, then so is ζ^{-1} .
15. Prove that if ζ is a primitive n -th root of unity, then $1, \zeta, \zeta^2, \dots, \zeta^{n-1}$ are all distinct.
16. Prove that if n is a positive odd integer and ζ is a primitive n -th root of unity, then so is ζ^2 . Is this also true for even n ? Justify your answer.
17. Prove Corollary 2.17.
18. Prove that if $\zeta \neq 1$ is any root of unity, then $1 + \zeta + \zeta^2 + \dots + \zeta^{\text{o}(\zeta)-1} = 0$.
19. Prove that if ζ is any root of unity, then $1 \cdot \zeta \cdot \zeta^2 \cdots \text{o}(\zeta)^{-1} = (-1)^{\text{o}(\zeta)-1}$.
20. Prove that if p is a prime number, then every imaginary p -th root of unity is necessarily a primitive p -th root of unity.
21. Prove that if $n > 4$ is not a prime, then at least three of the n -th roots of unity are not primitive n -th roots.
22. Prove that if $\zeta \in \sqrt[6]{1}$ and $\zeta \in \sqrt[22]{1}$, then $\zeta = \pm 1$.

2.6 The Existence of Complex Numbers

This section is devoted to the construction of a number system whose ontological credentials are impeccable and which is indistinguishable from the complex number system. An alternate proof of the existence of complex numbers is offered in Section 10.3 in a much wider and more useful setting.

We begin by defining a *Cartesian number* as an ordered pair (a, b) of real numbers. The two Cartesian numbers $z = (a, b)$ and $w = (c, d)$ are considered to be the same, or

equal, if and only if $a = c$ and $b = d$. Thus, $(2, 3) \neq (3, 2)$ and $(2^2, 3^3) = (4, 27)$. These Cartesian numbers can be thought of either as pairs of real numbers or as points of the Cartesian plane.

The addition and multiplication of Cartesian numbers are defined as follows:

$$\begin{aligned}(a, b) + (c, d) &= (a + c, b + d), \\ (a, b) \cdot (c, d) &= (ac - bd, ad + bc).\end{aligned}$$

These definitions are motivated by the fact that the Cartesian number (a, b) is supposed to be a logical construct that mimics the behavior of the intuitive quantity $a + bi$. Thus the addition and multiplication of Cartesian numbers mimic the facts that $(a + bi) + (c + di) = (a + c) + (b + d)i$ and $(a + bi)(c + di) = (ac - bd) + (ad + bc)i$.

This, of course, is only the motivation for these definitions. From the purely logical stance, these definitions need no justification. The addition and multiplication of Cartesian numbers can be demonstrated to possess all the usual properties that they have in the context of real numbers (these well-known properties will be formalized later in Section 6.1). Thus, the addition of Cartesian numbers is commutative because it inherits this property from the addition of real numbers in the following manner:

$$(a, b) + (c, d) = (a + c, b + d) = (c + a, d + b) = (c, d) + (a, b).$$

Similarly, the multiplication of Cartesian numbers is associative because

$$\begin{aligned}[(a, b) \cdot (c, d)] \cdot (e, f) &= (ac - bd, ad + bc) \cdot (e, f) \\ &= [(ac - bd)e - (ad + bc)f, (ac - bd)f + (ad + bc)e] \\ &= (ace - bde - adf - bcf, acf - bdf + ade + bce) \\ &= (ace - adf - bcf - bde, acf + ade + bce - bdf) \\ &= [a(ce - df) - b(cf + de), a(cf + de) + b(ce - df)] \\ &= (a, b) \cdot (ce - df, cf + de) \\ &= (a, b) \cdot [(c, d) \cdot (e, f)].\end{aligned}$$

The remaining proofs of the commutativity, associativity, and distributivity of the addition and multiplication of Cartesian numbers are relegated to Exercises 2.6.1 to 2.6.3. We

define the Cartesian zero to be the pair $(0, 0)$ and denote it by 0_c . It is clear that if $z = (a, b)$ is any Cartesian number, then

$$z + 0_c = (a, b) + (0, 0) = (a, b) = z.$$

We define the Cartesian unity to be the pair $(1, 0)$ and denote it by 1_c . Note that for any Cartesian number $z = (a, b)$,

$$z \cdot 1_c = (a, b) \cdot (1, 0) = (a \cdot 1 - b \cdot 0, a \cdot 0 + b \cdot 1) = (a, b) = z.$$

Finally, we address the existence of inverses. It is clear that $(-a, -b)$ is the additive inverse of (a, b) in the sense that

$$(a, b) + (-a, -b) = 0_c.$$

If $(a, b) \neq 0_c$, then $a^2 + b^2 \neq 0$ and so

$$(c, d) = \left(\frac{a}{a^2 + b^2}, \frac{-b}{a^2 + b^2} \right)$$

is a well-defined Cartesian number and it can be verified (Exercise 2.6.4) that (c, d) is the multiplicative inverse of (a, b) in the sense that $(a, b) \cdot (c, d) = 1_c$.

The foregoing discussion establishes that the set of Cartesian numbers together with the operations of addition and multiplication constitutes a bona fide number system. We now demonstrate that these numbers are just the complex numbers in disguise by identifying amongst them a copy of the real number system together with a quantity that behaves just as the imaginary number i is expected to behave. Observe that

$$(a, 0) + (c, 0) = (a + c, 0 + 0) = (a + c, 0)$$

and

$$(a, 0) \cdot (b, 0) = (a \cdot b - 0 \cdot 0, a \cdot 0 + b \cdot 0) = (ab, 0).$$

In other words, the Cartesian number $(a, 0)$ behaves with respect to Cartesian addition and multiplication just as the real number a behaves with respect to real addition and multiplication. Thus the set of Cartesian numbers whose second coordinate is 0 is indistinguishable from the real number system.

Let i_c denote the Cartesian number $(0, 1)$. Note that

$$i_c^2 = (0, 1) \cdot (0, 1) = (0 \cdot 0 - 1 \cdot 1, 0 \cdot 1 + 1 \cdot 0) = (-1, 0)$$

and $(-1, 0)$ is the Cartesian number that corresponds to the real number -1 . Moreover, if z is any Cartesian number (a, b) , then

$$\begin{aligned} (a, 0) + (b, 0)i_c &= (a, 0) + (b, 0) \cdot (0, 1) = (a, 0) + (b \cdot 0 - 0 \cdot 1, b \cdot 1 + 0 \cdot 0) \\ &= (a, 0) + (0, b) = (a, b) = z. \end{aligned}$$

This, of course, is the Cartesian analog of the fact that the arbitrary complex number z can be written in the form $a + bi$ where a and b are real numbers and i is a square root of -1 . Thus, we have shown that the Cartesian numbers, whose existence is justified by definition, behave just like the complex numbers.

Exercises 2.6

1. Prove that the addition of Cartesian numbers is associative.
2. Prove that the multiplication of Cartesian numbers is commutative.
3. Prove the distributivity of Cartesian numbers. I.e., prove that

$$(a, b)[(c, d) + (e, f)] = (a, b)(c, d) + (a, b)(e, f).$$

4. Prove that if $(a, b) \neq 0_c$, then

$$(a, b) \cdot \left(\frac{a}{a^2 + b^2}, \frac{-b}{a^2 + b^2} \right) = 1_c.$$

5. Prove that if $(a, b) \neq 0_c$ and $(a, b) \cdot (x, y) = (a, b)$ for some (x, y) , then $(x, y) = 1_c$.
6. Prove that $(a, b) \cdot 0_c = 0_c$.
7. Prove that if $(a, b) \neq 0_c$ and $(a, b) \cdot (c, d) = 0_c$, then $(c, d) = 0_c$.

Chapter Summary

This chapter began with an informal definition of the complex numbers and went on to a discussion of the four arithmetical operations and the extraction of roots in this new context, special emphasis being given to the roots of unity. These operations were used to give a formal definition of the concept of solvability by radicals. The geometry of the roots of unity was then used to prove the ruler-and-compass constructibility of the regular 17-sided polygon. This application relied on some surprising subtleties inherent in the powers of the roots of unity. The related notion of the order of a root of unity and some of its properties were expounded in the next section. Finally, a rationale justifying the existence of the so-called imaginary numbers was offered in the last section.

Chapter Review Exercises

Mark the following true or false.

1. The sum of two complex numbers is a complex number.
2. The sum of two imaginary numbers is never a real number.
3. The numbers 0, $1 + i$, $2 - i$, and i form a parallelogram.
4. $\arg[(1 + 2i)(1 - i)] = \arg(1 - i) + \arg(1 + 2i)$.
5. $|(1 + 2i)^{123}| = |(1 + 2i)|^{123}$.
6. The number 1 has twenty 20-th roots.
7. The number 1 has nineteen 20-th roots of order 20.
8. The elements of $\sqrt[7]{1}$ form a regular heptagon.
9. If $a \neq 0$, then the solutions of $ax^2 + bx + c = 0$ have a degree 2 algebraic expression in $\{a, b, c\}$.
10. The solutions of $3x^2 - 7x + 11 = 0$ are constructible.
11. The regular pentagon and the regular 17-sided polygon are the only regular polygons that are constructible.
12. The order of every element of $\sqrt[24]{1}$ is a divisor of 72.
13. If α is a primitive 7th root of unity, then α , α^2 , α^3 , and α^4 are all distinct.

New Terms

- algebraic expression, 23
- algebraic solution, 23
- algebraically resolvable, 24
- argument, 11
- argument principle, 14
- Cartesian number, 38
- Cartesian representation, 10
- complex number, 9
- composite, 37
- conjugate, 14
- constructible, 26
- divisor, 37
- doubling of a cube, 34
- imaginary number, 10
- modulus, 11
- multiple, 37
- order of a root of unity, 36
- polar form, 12
- prime, 37
- roots of unity, 18
- solvable by radicals, 24
- transcendental numbers, 34
- trisection of an angle, 34
- unit segment, 26

Supplementary Exercises

1. Write a computer script that will verify that the regular 257-sided polygon is ruler-and-compass constructible.
2. Write a computer script that will compute the n -th roots of any complex number.
3. Is there an analog of Exercise 2.2.30 for squares and other regular polygons?
4. Investigate the number system obtained by replacing the multiplication of Cartesian numbers with

$$(a, b) \cdot (c, d) = (ac + bd, ad + bc).$$

Which numbers have multiplicative inverses? What are the roots of unity like?

5. Make up your own number system and investigate it.

Chapter 3



SOLUTIONS OF EQUATIONS

IT IS OUR PURPOSE here to discuss solutions of equations from several different points of view. We will touch on the issues of existence of solutions, existence of formulas, solvability by radicals, and computation of solutions.

3.1 The Cubic Formula

We are now in position to present the modern version of the Ferro-Tartaglia-Cardano solution to the general cubic equation.

Theorem 3.1 Every cubic equation is solvable by radicals.

Proof. For the sake of simplification, we shall assume that the cubic equation we wish to solve has the form

$$x^3 + ax^2 + bx + c = 0. \quad (3.2)$$

It is clear that every cubic equation can be reduced to this form. Next, the problem is further simplified by transforming it to a form that is free of the x^2 term. This is accomplished by a transformation of the type $x = \alpha + y$, where the value of α will shortly be specified. Substituting $x = \alpha + y$ into Equation 3.2 we get

$$(\alpha + y)^3 + a(\alpha + y)^2 + b(\alpha + y) + c = 0$$

or

$$y^3 + (3\alpha + a)y^2 + (3\alpha^2 + 2a\alpha + b)y + (\alpha^3 + a\alpha^2 + b\alpha + c) = 0.$$

The choice of $\alpha = -a/3$ will clearly make the above coefficient of y^2 vanish. The equation now reduces to

$$y^3 + py + q = 0 \quad (3.3)$$

where

$$p = \frac{3b - a^2}{3} \quad \text{and} \quad q = \frac{27c + 2a^3 - 9ab}{27}.$$

To solve the reduced cubic of Equation 3.3 we shall make another transformation:

$$y = z + \frac{\beta}{z}$$

where the value of β will be chosen so that the resulting equation is solvable. When this value of y is substituted into Equation 3.3, we obtain

$$\left(z + \frac{\beta}{z}\right)^3 + p\left(z + \frac{\beta}{z}\right) + q = 0,$$

or

$$z^3 + (3\beta + p)z + q + \frac{3\beta^2 + \beta p}{z} + \frac{\beta^3}{z^3} = 0.$$

Miraculously, the choice of $\beta = -p/3$ causes the coefficients of both z and $1/z$ to vanish, leaving us with the equation

$$z^3 + q - \frac{p^3}{27z^3} = 0 \quad \text{or} \quad z^6 + qz^3 - \frac{p^3}{27} = 0. \quad (3.4)$$

This is a quadratic equation in z^3 with solutions

$$z^3 = \frac{-q \pm \sqrt{q^2 + \frac{4p^3}{27}}}{2}. \quad (3.5)$$

It is clear that z has an algebraic expression in p and q . Hence, if $z \neq 0$, each of the terms of the sequence p, q, z, y, x has an algebraic expression in $\{a, b, c\}$. If $z = 0$, then $y = z + \beta/z$ fails to be an algebraic expression. However, in that case, by Exercise 3.1.10, $x^3 + ax^2 + bx + c = (x + a/3)^3 + c - a^3/27$ and the corresponding equation is clearly algebraically resolvable. ■

The proof of Theorem 3.1 yields six possible values for z from which one can obtain six possible values for x . We shall later see (Proposition 6.8) that a cubic equation can have at most three roots, and we now set out to select three of the above six values that always yield all the solutions of the given cubic equation. The proof that these are the “correct” three values is deferred to Section 6.4. If all of the values of z that arise from

Equation 3.5 are 0, then (Exercise 3.1.10)

$$x^3 + ax^2 + bx + c = \left(x + \frac{a}{3}\right)^3$$

and so $x = -a/3$ is the triple solution of Equation 3.2. Otherwise, let z_1 be any nonzero value of z obtained from Equation 3.5 and set $z_2 = \omega z_1$ and $z_3 = \omega^2 z_1$. For each $i = 1, 2, 3$ set

$$y_i = z_i - \frac{p}{3z_i} \quad \text{and} \quad x_i = y_i - \frac{a}{3}. \quad (3.6)$$

Then $\{x_1, x_2, x_3\}$ is the complete solution set to Equation 3.2.

Consider the cubic equation $x^3 - 3x + 2 = 0$. Here $p = -3$, $q = 2$, and

$$\frac{-q \pm \sqrt{q^2 + \frac{4p^3}{27}}}{2} = \frac{-2 \pm \sqrt{4 + \frac{4(-3)^3}{27}}}{2} = -1,$$

and so we can choose $z_1 = -1$. This gives

$$\begin{aligned} x_1 = y_1 &= -1 - \frac{-3}{3(-1)} = -2, \\ x_2 = y_2 &= \omega(-1) - \frac{-3}{3\omega(-1)} = -\omega - \omega^2 = 1, \\ x_3 = y_3 &= \omega^2(-1) - \frac{-3}{3\omega^2(-1)} = -\omega^2 - \omega = 1. \end{aligned}$$

We next turn to an example with complex coefficients—the equation $x^3 + 3ix - (1 + i) = 0$. Here $p = 3i$, $q = -(1 + i)$, and

$$\begin{aligned} \frac{-q + \sqrt{q^2 + \frac{4p^3}{27}}}{2} &= \frac{1 + i + \sqrt{(1 + i)^2 + \frac{4(3i)^3}{27}}}{2} \\ &= \frac{1 + i + \sqrt{1 + 2i - 1 - 4i}}{2} \\ &= \frac{1 + i + 1 - i}{2} \\ &= 1. \end{aligned}$$

Consequently, we may choose $z_1 = 1$ and so

$$\begin{aligned}x_1 &= y_1 = 1 - \frac{3i}{3 \cdot 1} = 1 - i, \\x_2 &= y_2 = 1 \cdot \omega - \frac{3i}{3 \cdot \omega \cdot 1} = \omega - \omega^2 i, \\x_3 &= y_3 = 1 \cdot \omega^2 - \frac{3i}{3 \cdot \omega^2 \cdot 1} = \omega^2 - \omega i.\end{aligned}$$

Finally, we examine an equation in which the coefficient of x^2 is not zero, namely, the equation $x^3 - 6x^2 - 4 = 0$. Here $a = -6$, $b = 0$, and $c = -4$, so that $p = -12$ and $q = -20$. Substitution of these values into Equation 3.5 yields $z^3 = 10 \pm 6$. Choosing $z_1 = \sqrt[3]{4}$ we then obtain

$$\begin{aligned}x_1 &= y_1 - \frac{a}{3} = \sqrt[3]{4} - \frac{-12}{3\sqrt[3]{4}} - \frac{-6}{3} = 2 + \sqrt[3]{4} + \sqrt[3]{16}, \\x_2 &= y_2 - \frac{a}{3} = \sqrt[3]{4}\omega - \frac{-12}{3\sqrt[3]{4}\omega} - \frac{-6}{3} = 2 + \sqrt[3]{4}\omega + \sqrt[3]{16}\omega^2, \\x_3 &= y_3 - \frac{a}{3} = \sqrt[3]{4}\omega^2 - \frac{-12}{3\sqrt[3]{4}\omega^2} - \frac{-6}{3} = 2 + \sqrt[3]{4}\omega^2 + \sqrt[3]{16}\omega.\end{aligned}$$

In conclusion we note that while the solution of the cubic equation given in Equation 3.6 above is formally different from that which appears as Equation 1.5, it is not too difficult to show that this latter expression actually equals one of the three roots that are given by Equation 3.6. The details are relegated to Exercise 3.1.17.

Exercises 3.1

Use the cubic formula to find all the complex roots of each of the equations in Exercises 3.1.1 to 3.1.9.

1. $x^3 + 9x - 6 = 0$
2. $x^3 + 12x - 12 = 0$
3. $x^3 + 18x - 30 = 0$
4. $x^3 - 15x - 126 = 0$
5. $x^3 + 3x^2 + 9x + 5 = 0$
6. $x^3 - 6x^2 + 24x - 44 = 0$
7. $x^3 + 3x^2 - 3x + 2i - 5 = 0$
8. $x^3 + 6ix - 1 - 8i = 0$ (Hint: $\sqrt{-63 - 16i} = \pm(1 - 8i)$.)
9. $2x^3 + 3x^2 + 3x + 1 = 0$

10. Prove that if all the values of z in Equation 3.5 are 0, then $p = q = 0$ and $x^3 + ax^2 + bx + c = (x + a/3)^3$.

Use Exercise 3.1.10 to solve Exercises 3.1.11 and 3.1.12.

11. $x^3 + 3(1+i)x^2 + 6ix - 2(1-i) = 0$

12. $x^3 + 6ix^2 - 12x - 8i = 0$

If x_1 , x_2 , and x_3 are the solutions of the cubic equation $x^3 + ax^2 + bx + c = 0$, then the quantity $[(x_1 - x_2)(x_2 - x_3)(x_3 - x_1)]^2$ is called the *discriminant* of the given equation.

13. Prove that the discriminant of the equation $y^3 + py + q = 0$ is $-4p^3 - 27q^2$.
14. Prove that the discriminant of the equation $x^3 + ax^2 + bx + c = 0$ is $18abc - 4a^3c + a^2b^2 - 4b^3 - 27c^2$.
15. Prove that a cubic equation with real coefficients has three distinct roots if its discriminant D is positive, a single real root if D is negative, and at least two equal real roots if D is zero.
16. Show that if a is real, then the equation $x^3 + ax + 2 = 0$ has three real roots if and only if $a \leq -3$.
17. Show that the root of the cubic equation given in Equation 1.5 is also one of those given by Equation 3.6 of this section.

3.2 Solvability by Radicals II

Whereas thousands of years elapsed between the solutions of the quadratic and the general cubic equation, only a few more years passed before Cardano assigned the problem of finding a formula for the general *fourth-degree equation*, or *quartic equation*, to his disciple Lodovico Ferrari and the latter succeeded in solving it. Subsequently, several other mathematicians presented their own solutions of the quartic. From our perspective, the most significant of these is Lagrange's solution which is presented in detail in Section 6.5.

Quite naturally, mathematicians next turned their attention to the general *quintic equation*, or *fifth-degree equation*. This equation, however, presented new difficulties, and no substantial progress was made for another 250 years. During the second half of the eighteenth century some mathematicians recognized the fact that even the innocent looking equation

$$x^n - 1 = 0 \quad (3.7)$$

presented them with challenges. They were aware that one of the solutions of the equation $x^{11} - 1 = 0$ is the complex number

$$\cos \frac{2\pi}{11} + i \sin \frac{2\pi}{11},$$

and consequently this number can be expressed in terms of a radical of the 11th order (namely $\sqrt[11]{1}$). However, in 1771 Vandermonde and Lagrange showed that the same number could also be expressed in terms of radicals of the 2nd and 5th orders, thus uncovering some surprising relationships between radicals of the 2nd, 5th, and 11th orders.

Equation 3.7 came to be known as the *cyclotomic equation* because its zeros, as stated in Proposition 2.6, lie on a common circle. In 1801, in the concluding chapter of his *Disquisitiones Arithmeticae*, Gauss delved into the subtleties of the radicals that are required for the solution of the cyclotomic equation. The proof of Theorem 2.15 was meant to provide a sample of the algebraic manipulations that this task required. The very special nature of the cyclotomic equation notwithstanding, this work turned out to be very seminal. Évariste Galois, who completely settled the issue of algebraic resolvability of equations some 30 years later, made repeated references to Gauss's techniques and results in explaining the motivation for his own work.

3.3 Other Types of Solutions

While this book's main theme is the issue of solvability of polynomial equations by means of algebraic operations, it might be pedagogically advantageous at this point to discuss some other senses in which an equation could be solvable. Consider the equation

$$x^5 - 6x + 3 = 0. \tag{3.8}$$

If we set $f(x) = x^5 - 6x + 3$, then

$$\lim_{x \rightarrow \infty} f(x) = \infty \quad \text{and} \quad \lim_{x \rightarrow -\infty} f(x) = -\infty.$$

Since $f(x)$ is a continuous function, it follows that its graph must cross the x -axis at some point and that point clearly yields a solution to Equation 3.8. This argument generalizes easily to the following:

Theorem 3.9 If n is an odd positive integer, and a_1, a_2, \dots, a_n are real numbers, then the equation

$$x^n + a_1 x^{n-1} + a_2 x^{n-2} + \cdots + a_{n-1} x + a_n = 0$$

has a real solution.

Thus, it is possible to argue the existence of a solution to an equation without having any information at all about the value of the solution. In fact, the mathematicians of the seventeenth and eighteenth centuries became convinced of the validity of the following sweeping statement:

Theorem 3.10 Every equation of the form $x^n + a_1 x^{n-1} + a_2 x^{n-2} + \cdots + a_{n-1} x + a_n = 0$ has a solution.

The solution whose existence is guaranteed here may be complex, but it always exists. The first valid proof of this fact was provided by Gauss in 1796 and it eventually became known as the *Fundamental Theorem of Algebra*. Neither this proof, nor any of Gauss's several subsequent alternate proofs provided information about the actual value of the solution; they were concerned only with its existence. In 1826, Niels Abel proved that no analogs of the Ferro-Tartaglia-Cardano cubic formula could exist for the general fifth-degree equation,

$$x^5 + ax^4 + bx^3 + cx^2 + dx + e = 0.$$

Shortly thereafter Évariste Galois constructed a general theory for determining just which general equations have formulas and which specific equations are solvable by radicals. Using that theory it is possible, for example, to show that the roots of Equation 3.8, the existence of at least one of which is easily demonstrated, do not have algebraic expressions in the integers.

There is yet another aspect to solving equations that is essentially different from both the issue of existence and from that of expressibility by algebraic operations, and that is the problem of evaluating a root and writing it down as, say, a decimal number. Just because it is known that a certain number is an algebraic expression in the integers does not mean that we have any idea of its size. The root of Equation 3.8 whose existence was proven by a theoretical argument, of course, also suffers from the same lack of precision.

There are numerical methods for finding roots of equations which are much more practical than even the Ferro-Tartaglia-Cardano formula. The best known of these, the *Newton-Raphson method*, is usually taught in the first semester of calculus but will never-

theless be reviewed here because we feel that this will bring about a better understanding of the difference between solvability of equations in general and solvability of equations by radicals.

Loosely speaking, the Newton-Raphson method, when applied to an equation of the form $f(x) = 0$, says that if x_n is an estimate for a solution, then

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \quad (3.11)$$

where $f'(x)$ denotes the derivative of $f(x)$ with respect to x , is a better estimate for the same solution. Consider Equation 3.8. With $f(x) = x^5 - 6x + 3$, we find that $f(0) = 3$ and $f(1) = -2$. Thus this equation must have a solution between 0 and 1, and we begin with $x_1 = 0$ as our first estimate. Since $f'(x) = 5x^4 - 6$, the Newton-Raphson method yields the following successive estimates (correct to four decimal places) for the solution:

$$\begin{aligned} x_2 &= 0 - \frac{0^5 - 6 \cdot 0 + 3}{5 \cdot 0^4 - 6} = 0.5, \\ x_3 &= 0.5 - \frac{(0.5)^5 - 6 \cdot 0.5 + 3}{5 \cdot (0.5)^4 - 6} = 0.5055, \\ x_4 &= 0.5055 - \frac{(0.5055)^5 - 6 \cdot 0.5055 + 3}{5 \cdot (0.5055)^4 - 6} = 0.5055. \end{aligned}$$

Once two successive estimates are equal to each other, there will, of course, be no further improvement in the estimates unless the number of decimal places is increased. It is easily verified that

$$(0.5055)^5 - 6(0.5055) + 3 = 0.000006981 \dots$$

and that the solution, correct to six decimal places, is 0.550501.

This is not the book in which to either discuss the subtleties or attempt a rigorous proof of this marvelous technique. The underlying idea, though, is surprisingly simple. Suppose we have arrived at the estimate of x_n for the solution r of the equation $f(x) = 0$ (Figure 3.1). If $P = (x_n, f(x_n))$ is the point of the graph of $y = f(x)$ that lies directly above the x -axis point $(x_n, 0)$, then the equation of the line through P and tangent to this graph is given by the point-slope formula as

$$y - f(x_n) = f'(x_n)(x - x_n). \quad (3.12)$$

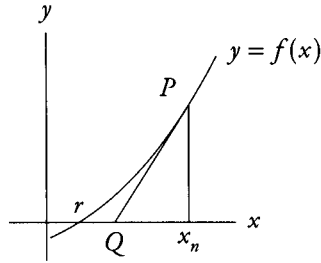


Figure 3.1 The Newton-Raphson method

Note that the diagram suggests that the x -intercept Q of the tangent line at P lies closer to our goal of $R = (r, 0)$ than P . We therefore choose the abscissa of this intersection as the next estimate x_{n+1} . Since Q is the x -intercept of the tangent line at P , the value of x_{n+1} is easily obtained by substituting $(x_{n+1}, 0)$ for (x, y) in Equation 3.12. This yields

$$0 - f(x_n) = f'(x_n)(x_{n+1} - x_n).$$

When this equation is solved for x_{n+1} , we obtain the Newton-Raphson recursion of Equation 3.11.

The Newton-Raphson method we presented here is not perfect. There are cases wherein this method will miss some rather obvious answers. Nevertheless, when this method does locate a root, the root will be correct.

Exercises 3.3

Using the Newton-Raphson method, find at least one real root of each of the equations in Exercises 3.3.1 to 3.3.8 (use four decimal places).

1. $x^3 - 3x^2 + 18x + 12 = 0$
2. $2x^3 - 5x^2 + x - 10 = 0$
3. $x^4 - 100x - 85 = 0$
4. $x^4 + x^3 - x^2 - 4x - 13 = 0$
5. $x^5 - 30x + 17 = 0$
6. $x^7 + x^5 - 3 = 0$
7. $\sin x = x + 1$ (Set $f(x) = \sin x - x - 1$.)
8. $\ln x = \sin x$

Using the Newton-Raphson method (to four decimal places), estimate the real roots in Exercises 3.3.9 to 3.3.12.

9. $\sqrt[3]{10}$ (Solve the equation $x^3 - 10 = 0$.)

10. $\sqrt[4]{100}$

11. $\sqrt[5]{10000}$

12. $\sqrt[8]{54,321}$

For each of the equations in Exercises 3.3.13 to 3.3.18, explain why it does or does not have real solutions.

13. $x^7 + x^5 - 3 = 0$

16. $x^{16} - 4x^8 + 5 = 0$

14. $x^9 + 101x^8 + 93x^6 + 41x^4 + 1 = 0$

17. $\ln x = \tan x$

15. $x^8 - 2x^4 + 1 = 0$

18. $x^3 = \ln x$

19. Let a , b , and c be real numbers. Prove that if $a^2 \leq 3b$, then the equation $x^3 + ax^2 + bx + c = 0$ has exactly one real root.

Chapter Summary

We began by showing that every cubic equation is solvable by radicals. While the issue of solvability of equations by radicals eventually led to the creation of modern algebra, it was pointed out in this chapter that there are other, no less significant, aspects to the solvability of equations. The question of the existence of roots can be treated without regard to explicit derivations. Thus, it is known that every equation with either complex or real coefficients has at least one, possibly complex, solution. Ad hoc arguments can be given for the existence of such roots that provide no information about its value. Finally, the Newton-Raphson method was informally discussed to show how roots can be found without the use of radicals.

Chapter Review Exercises

Mark the following true or false.

1. Every cubic equation has three distinct roots.
2. The equation $x^3 + ax^2 + bx + c = 0$ has at least one real root.
3. Every cubic equation has at least one imaginary solution.
4. Every equation can be solved by the Newton-Raphson method.
5. Every root of the equation $x^{23} - 1 = 0$ has an algebraic expression in the integers.

New Terms

cyclotomic equation, 50

fifth-degree equation, 49

fourth-degree equation, 49

Fundamental Theorem of Algebra, 51

Newton-Raphson method, 51

quartic equation, 49

quintic equation, 49

Supplementary Exercises

1. Implement the cubic formula on a computer.
2. Implement the Newton-Raphson method on a calculator or a computer.

Chapter 4



MODULAR ARITHMETIC

THIS CHAPTER introduces some new number systems that are suggested by the properties of the exponents of the complex roots of unity. These number systems bear striking similarities to the more traditional rational and real numbers but also differ from them in crucial ways.

4.1 Modular Addition, Subtraction, and Multiplication

One of the steps in Gauss's proof of the ruler-and-compass constructibility of the regular 17-sided polygon called for the verification of the identity

$$(\zeta + \zeta^{16})(\zeta^{13} + \zeta^4) = \zeta^{14} + \zeta^5 + \zeta^{29} + \zeta^{20} = \zeta^{14} + \zeta^5 + \zeta^{12} + \zeta^3.$$

In his writings Gauss used an abbreviation that replaced each ζ^k with the symbol $[k]$. Since $\zeta^{k+17} = \zeta^k$, it follows that, in Gauss's notation, $[k+17] = [k]$ for each integer k . This is an example of *modular arithmetic*. For any positive integer n , the two integers a and b are said to be *congruent modulo n* , and we write

$$a \equiv b \pmod{n},$$

whenever n is a divisor of $a - b$. This, by Corollary 2.17, is tantamount to saying that $\zeta^a = \zeta^b$ where ζ is any primitive n -th root of 1. Thus, $7 \equiv 3 \pmod{4}$, $2 \equiv 14 \pmod{6}$, and $-3 \equiv 35 \pmod{19}$. Note that if $a \equiv a' \pmod{n}$ and $b \equiv b' \pmod{n}$, and if ζ is as above, then

$$\zeta^{a+b} = \zeta^a \zeta^b = \zeta^{a'} \zeta^{b'} = \zeta^{a'+b'}$$

and

$$\zeta^{ab} = (\zeta^a)^b = (\zeta^{a'})^b = (\zeta^b)^{a'} = (\zeta^{b'})^{a'} = \zeta^{a'b'}$$

$(\mathbb{Z}_4, +)$	0	1	2	3
0	0	1	2	3
1	1	2	3	0
2	2	3	0	1
3	3	0	1	2

(\mathbb{Z}_4, \cdot)	0	1	2	3
0	0	0	0	0
1	0	1	2	3
2	0	2	0	2
3	0	3	2	1

Table 4.1 Arithmetic modulo 4

$(\mathbb{Z}_5, +)$	0	1	2	3	4
0	0	1	2	3	4
1	1	2	3	4	0
2	2	3	4	0	1
3	3	4	0	1	2
4	4	0	1	2	3

(\mathbb{Z}_5, \cdot)	0	1	2	3	4
0	0	0	0	0	2
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
3	0	4	3	2	1

Table 4.2 Arithmetic modulo 5

and hence, $a + b \equiv a' + b' \pmod{n}$ and $ab \equiv a'b' \pmod{n}$.

It follows from this that when performing arithmetic modulo n , it suffices to consider the applications of the arithmetic operations to the integers $0, 1, 2, \dots, n-1$ alone. Tables 4.1 to 4.4 contain such abbreviated addition and multiplication tables for $n = 4, 5, 6$, and 7 . Motivated by the standard denotation of the set of integers by \mathbb{Z} , arithmetic modulo n , when restricted to the set $\{0, 1, \dots, n-1\}$, is denoted by \mathbb{Z}_n . An alternate, and more formal, definition of modular arithmetic is offered in Section 10.1 following Corollary 10.7.

The commutativity, associativity, and distributivity of modular addition and multiplication are consequences of the following identities: $\zeta^{a+b} = \zeta^{b+a}$, $\zeta^{(a+b)+c} = \zeta^{a+(b+c)}$, $\zeta^{ab} = \zeta^{ba}$, $\zeta^{(ab)c} = \zeta^{a(bc)}$, and $\zeta^{a(b+c)} = \zeta^{ab+ac}$. It is also clear that $a \cdot 0 \equiv 0 \cdot a \equiv 0 \pmod{n}$ and $a \cdot 1 \equiv 1 \cdot a \equiv a \pmod{n}$ since $\zeta^{a \cdot 0} = \zeta^{0 \cdot a} = \zeta^0$ and $\zeta^{a \cdot 1} = \zeta^{1 \cdot a} = \zeta^a$.

Thus, the operations of addition and multiplication possess the same desirable properties relative to modular equivalence that they have when applied to the integers in the context of conventional equality.

$(\mathbb{Z}_6, +)$	0	1	2	3	4	5	(\mathbb{Z}_6, \cdot)	0	1	2	3	4	5
0	0	1	2	3	4	5	0	0	0	0	0	0	0
1	1	2	3	4	5	0	1	0	1	2	3	4	5
2	2	3	4	5	0	1	2	0	2	4	0	2	4
3	3	4	5	0	1	2	3	0	3	0	3	0	3
4	4	5	0	1	2	3	4	0	4	2	0	4	2
5	5	0	1	2	3	4	5	0	5	4	3	2	1

Table 4.3 Arithmetic modulo 6

$(\mathbb{Z}_7, +)$	0	1	2	3	4	5	6
0	0	1	2	3	4	5	6
1	1	2	3	4	5	6	0
2	2	3	4	5	6	0	1
3	3	4	5	6	0	1	2
4	4	5	6	0	1	2	3
5	5	6	0	1	2	3	4
6	6	0	1	2	3	4	5

(\mathbb{Z}_7, \cdot)	0	1	2	3	4	5	6
0	0	0	0	0	0	0	0
1	0	1	2	3	4	5	6
2	0	2	4	6	1	3	5
3	0	3	6	2	5	1	4
4	0	4	1	5	2	6	3
5	0	5	3	1	6	4	2
6	0	6	5	4	3	2	1

Table 4.4 Arithmetic modulo 7

If a is any nonzero element of \mathbb{Z}_n , then $n - a$ is also in \mathbb{Z}_n and

$$a + (n - a) = n \equiv 0 \pmod{n},$$

and so $n - a$ is the additive inverse $-a$ of a in \mathbb{Z}_n . Thus, 3 and 4 are each other's additive inverses in \mathbb{Z}_7 , and 3 and 5 are each other's additive inverses in \mathbb{Z}_8 . The additive inverse of 0 is, by definition, itself. The guaranteed existence of these additive inverses makes it possible to define the difference $a - b$ of two elements of \mathbb{Z}_n as the element $a + (-b)$. Thus, in \mathbb{Z}_{13} , $5 - 7 \equiv 5 + 6 \equiv 11$ and $5 - 2 \equiv 5 + 11 \equiv 3$. A moment's reflection will lead to the conclusion that whenever a and b are integers such that $0 \leq b \leq a \leq n - 1$, then $a - b$ is unambiguous, regardless of whether a and b are considered as conventional integers or as elements of \mathbb{Z}_n .

The issues of the existence of multiplicative inverses and the possibility of division in \mathbb{Z}_n are more subtle. There is no integer x such that $2x \equiv 1 \pmod{4}$ since in the (\mathbb{Z}_4, \cdot) multiplication table (Table 4.1) the row corresponding to 2 does not contain the entry 1. On the other hand, in the (\mathbb{Z}_5, \cdot) multiplication table (Table 4.2) all the rows but the first contain a 1, implying that every nonzero element of \mathbb{Z}_5 has a multiplicative inverse in \mathbb{Z}_5 . Thus, $2 \cdot 3 \equiv 1 \pmod{5}$ and $4 \cdot 4 \equiv 1 \pmod{5}$, meaning that 2 and 3 are each other's multiplicative inverses in \mathbb{Z}_5 whereas 4 is its own multiplicative inverse, a not so surprising fact, since $4 \equiv -1 \pmod{5}$.

A glance at the multiplication tables for \mathbb{Z}_6 and \mathbb{Z}_7 makes it clear that the situation here is entirely analogous to the above. The integer 2 has no multiplicative inverse in \mathbb{Z}_6 whereas every nonzero element of \mathbb{Z}_7 does have such an inverse. It will be seen in the next section that \mathbb{Z}_n possesses all the requisite multiplicative inverses if and only if n is a prime number.

A word of caution is in order here. While it is true that $7 \equiv 2 \pmod{5}$, it does not follow from this that $x^7 \equiv x^2 \pmod{5}$ since $2^7 = 128 \equiv 4 = 2^2 \pmod{5}$.

Even very complicated looking equations can be easily solved in modular arithmetic by the naive method of substitution. In \mathbb{Z}_5 , the solution set of the equation

$$x^6 + 3x^4 + 2x + 4 = 0$$

is $\{1, 2\}$ because these are the only elements of \mathbb{Z}_5 whose substitution into the polynomial $x^6 + 3x^4 + 2x + 4$ yields $0 \pmod{5}$.

Exercises 4.1

Solve the equations in Exercises 4.1.1 to 4.1.5 in \mathbb{Z}_2 .

1. $x + 1 \equiv 0$ 2. $x^2 + 1 \equiv 0$ 3. $x^3 + x \equiv 0$
 4. $x^5 + x^2 + x + 1 \equiv 0$ 5. $x^{17} + x^5 + x^3 \equiv 0$

6. Solve the equations of Exercises 4.1.1 to 4.1.5 in \mathbb{Z}_3 .

7. Solve the equations of Exercises 4.1.1 to 4.1.5 in \mathbb{Z}_5 .

Solve the equations in Exercises 4.1.8 to 4.1.15 in \mathbb{Z}_7 .

8. $3x + 2 \equiv 1$ 10. $x^3 + x^2 + 4x + 1 \equiv 0$
 9. $5x^2 + 2x - 1 \equiv 0$ 11. $4x - 3y \equiv 1$ and $2x + 4y \equiv 3$
 12. $4x + 3y \equiv 1$ and $x + 6y \equiv 3$
 13. $x + 2y + 3z \equiv 1$, $3x + y + 2z \equiv 2$, and $2x + 3y + z \equiv 3$
 14. $x^{123,456} + x + 5 \equiv 0$ 15. $x^{54,321} + 2x^{5,432} + 3x + 1 \equiv 0$

16. Solve the equations in Exercises 4.1.11 to 4.1.13 in \mathbb{Z}_5 .

17. Solve the equations in Exercises 4.1.11 and 4.1.12 in \mathbb{Z}_{13} .

18. Solve the equations in Exercises 4.1.14 and 4.1.15 in \mathbb{Z}_5 .

19. Solve the equations in Exercises 4.1.14 and 4.1.15 in \mathbb{Z}_{13} .

20. Evaluate the (modulo n) sum of all the elements of \mathbb{Z}_n for each positive integer n .

21. Evaluate $1 + 3 + 5 + \cdots + 1,001$ in $\mathbb{Z}_{1,002}$.

22. Evaluate $1 + 4 + 7 + \cdots + 1,234$ in \mathbb{Z}_{432} .

23. Evaluate $1 + 2 + 4 + \cdots + 2^{63}$ in \mathbb{Z}_{11} .

24. Evaluate $1 + 3 + 9 + \cdots + 3^{99}$ in \mathbb{Z}_7 .

25. Prove that an integer is divisible by 3 if and only if the sum of its digits in base 10 notation is divisible by 3.

26. Prove that an integer is divisible by 9 if and only if the sum of its digits in base 10 notation is divisible by 9.

27. Let ζ be the first n -th root of unity, and let $\eta = \zeta^k$ be any other n -th root of unity. Prove that there exists an integer m such that $\zeta = \eta^m$ if and only if k has a multiplicative inverse in \mathbb{Z}_n .

4.2 The Euclidean Algorithm and Modular Inverses

It turns out that the best way to deal with the question of invertible elements in \mathbb{Z}_n involves the Euclidean algorithm for finding the greatest common divisor of two integers. This is a problem that was considered by many of the earliest mathematicians, including Euclid. An integer that is a divisor of both the integers m and n is said to be a *common divisor*. If such a common divisor of m and n has the additional property that it is divisible by all the common divisors of m and n , then it is the *greatest common divisor* (GCD) or *highest common factor* (HCF) of m and n and it is denoted by (m, n) . Thus, ± 1 , ± 2 , ± 3 , ± 4 , ± 6 , and ± 12 are all the common factors of 24 and 36, but their HCF is 12. Thus,

$$12 = (24, 36) = (-24, 36) = (-24, -36).$$

Note that since every integer divides 0, it follows that $(0, 0)$ does not exist. In Propositions 1 and 2 of Book VII of *The Elements* Euclid suggests the following method for finding the greatest common divisor of the two positive integers $m \geq n$. Suppose first that n is a divisor of m . Then it is clear that $(m, n) = n$. If n is not a divisor of m , then n and $m - n$ are positive integers such that $(m - n, n) = (m, n)$. The reason for this is that every common divisor of m and n is clearly also a common divisor of $m - n$ and n , and, vice versa, every common divisor of $m - n$ and n is also a divisor of $m = (m - n) + n$ and n . In other words, the set of common divisors of the pair $\{m - n, n\}$ is identical with the set of common divisors of the pair $\{m, n\}$. Consequently, the two pairs also have the same greatest common divisor.

This leads to the following derivation of the greatest common divisor of 481 and 74:

$$(481, 74) = (407, 74) = (333, 74) = (259, 74) = (185, 74) = (111, 74) = (37, 74) = 37.$$

It is clear that this procedure will always yield the greatest common divisor in a finite number of steps, and hence it does deserve to be called an algorithm. It is also clear that the repeated subtractions that lead from $(481, 74)$ to $(37, 74)$ could be abbreviated by observing that 37 is the remainder left by 481 when divided by 74. Thus, if p is greater than q and if p leaves remainder r when divided by q , we have $(p, q) = (r, q)$. This leads to a much faster algorithm for finding greatest common divisors.

To find the greatest common divisor of 2,227 and 12,707, we note that the remainder left by 12,707 upon division by 2,227 is 1,572 and hence $(12,707, 2,227)$ equals

(2,227, 1,572). Several applications of this reduction yield

$$(12,707, 2,227) = (1,572, 2,227) = (1,572, 655) = (262, 655) = (262, 131) = 131.$$

This highly efficient procedure for determining the greatest common divisor of two integers, commonly known as the *Euclidean algorithm*, has a very surprising range of nonobvious applications. Here, of course, we are interested in the question of which integers possess multiplicative inverses in modulo n arithmetic. Two numbers are said to be *relatively prime* whenever their greatest common divisor is 1. It is exactly those integers that are relatively prime to n that turn out to possess multiplicative inverses modulo n . The following proposition paves the way to proving this fact.

Proposition 4.1 Let m and n be two integers which are not both 0. Then $g = (m, n)$ exists, as do two integers A and B such that $(m, n) = Am + Bn$.

Proof. It suffices to prove this proposition when m and n are both positive (Exercise 4.2.25). Thus, we assume that $m \geq n \geq 0$ and proceed by induction on the number of steps (divisions) in the application of the Euclidean algorithm to m and n .

If the pair (m, n) is such that exactly one division is required by the Euclidean algorithm, that must be because m is a multiple of n , say $m = dn$ for some integer d . In that case $g = n$ and so $g = 0 \cdot m + 1 \cdot n$ is the required expression.

Assume next that $m \geq n$ are a fixed pair of positive integers the derivation of whose greatest common divisor requires $k > 0$ divisions, and that the proposition holds for all positive integers x and y the derivation of whose greatest common divisor requires $k - 1$ divisions. Let $q \geq 0$ and $r < n$ be the respective quotient and remainder of m when divided by n so that $m = qn + r$. By the abbreviated Euclidean algorithm, $(m, n) = (r, n)$. However, the derivation of (r, n) clearly requires one division less than the derivation of (m, n) . In other words, the derivation of (r, n) requires only $k - 1$ divisions, and so, by the induction hypothesis, there exist integers A' and B' such that

$$g = (m, n) = (r, n) = A'n + B'r.$$

However, by the definition of q and r , $m = qn + r$ so that

$$g = A'n + B'r = A'n + B'(m - qn) = B'm + (A' - B'q)n. \quad (4.2)$$

Hence,

$$A = B' \quad \text{and} \quad B = A' - B'q \quad (4.3)$$

are the required integers. \blacksquare

Corollary 4.4 If m is relatively prime to n , then m has a multiplicative inverse in \mathbb{Z}_n .

Proof. Let m and n be relatively prime, so that $(m, n) = 1$. By Proposition 4.1 there exist integers A and B such that $Am + Bn = 1$. Since $Bn \equiv 0 \pmod{n}$, it follows that $Am \equiv 1 \pmod{n}$, so that A is the multiplicative inverse of m in \mathbb{Z}_n . \blacksquare

Since

$$1 \cdot 1 \equiv 3 \cdot 3 \equiv 5 \cdot 5 \equiv 7 \cdot 7 \equiv 1 \pmod{8},$$

it follows that 1, 3, 5, and 7, which are all relatively prime to 8, are their own multiplicative inverses in \mathbb{Z}_8 . Similarly, since

$$1 \cdot 1 \equiv 2 \cdot 5 \equiv 4 \cdot 7 \equiv 8 \cdot 8 \equiv 1 \pmod{9},$$

it follows that the multiplicative inverses of 1, 2, 4, 5, 7, and 8 in \mathbb{Z}_9 are 1, 5, 7, 2, 4, and 8, respectively.

The converse of Corollary 4.4 also holds (Exercise 4.2.26). Whenever n is a prime number, each positive integer less than n is necessarily relatively prime to n , so that each of those has a multiplicative inverse in \mathbb{Z}_n . This fact is crucial to the subsequent development of this book and so we state it explicitly.

Proposition 4.5 If p is a prime number and m is an integer, $0 < m < p$, then m has a multiplicative inverse in \mathbb{Z}_p .

The list below indicates the multiplicative inverses of all the nonzero elements of \mathbb{Z}_2 , \mathbb{Z}_3 , \mathbb{Z}_5 , and \mathbb{Z}_7 :

$$1 \cdot 1 \equiv 1 \pmod{2},$$

$$1 \cdot 1 \equiv 2 \cdot 2 \equiv 1 \pmod{3},$$

$$1 \cdot 1 \equiv 2 \cdot 3 \equiv 4 \cdot 4 \equiv 1 \pmod{5},$$

$$1 \cdot 1 \equiv 2 \cdot 4 \equiv 3 \cdot 5 \equiv 6 \cdot 6 \equiv 1 \pmod{7}.$$

Let p be a fixed prime number and let m be any integer that is not divisible by p . Denote the multiplicative inverse of m by m^{-1} . We can then define

$$\frac{a}{b} \equiv a b^{-1} \pmod{p}$$

whenever $b \not\equiv 0 \pmod{p}$. Accordingly, $3/4 \equiv 3 \cdot 2 \equiv 6 \pmod{7}$. Thus, from the point of view of the four arithmetical operations, modulo p arithmetic (when p is a prime) is just as well behaved as the arithmetic of the rational numbers, or the real numbers, or the complex numbers. This observation will be formalized in Section 6.1.

The proof of Proposition 4.1 provides an effective method for finding a^{-1} whenever it exists in \mathbb{Z}_n . Thus, to find the multiplicative inverse of 37 in \mathbb{Z}_{201} we begin to answer this question by applying the Euclidean algorithm to 201 and 37. This gives

$$(201, 37) = (37, 16) = (16, 5) = (5, 1) = 1,$$

with

$$201 = 5 \cdot 37 + 16,$$

$$37 = 2 \cdot 16 + 5,$$

$$16 = 3 \cdot 5 + 1,$$

$$5 = 5 \cdot 1.$$

Now,

$$1 = 0 \cdot 5 + 1 \cdot 1,$$

and when $A' = 0$, $B' = 1$, $m = 16$, $n = 5$, and $q = 3$ are substituted in Equations 4.2 and 4.3, we get $A = 1$ and $B = -3$, so that

$$1 = 1 \cdot 16 + (-3) \cdot 5.$$

Again, when $A' = 1$, $B' = -3$, $m = 37$, $n = 16$, $q = 2$ are substituted in Equations 4.2 and 4.3, we get $A = -3$ and $B = 7$ so that

$$1 = (-3) \cdot 37 + 7 \cdot 16.$$

Finally, when $A' = -3$, $B' = 7$, $m = 201$, $n = 37$, and $q = 5$ are substituted in Equations 4.2 and 4.3, we get $A = 7$ and $B = -38$ so that

$$1 = 7 \cdot 201 + (-38) \cdot 37.$$

This means that $-38 \equiv 163$ is the multiplicative inverse of 37 in \mathbb{Z}_{201} .

This section's final propositions state some basic number theoretic facts that may seem rather obvious. It is greatly to Euclid's credit that he realized that these facts actually called for proofs. The proofs we offer boil down to a clever application of the Euclidean algorithm and are essentially the same as those which appear in Euclid's *The Elements*. The content of these propositions will be assumed by several subsequent proofs.

Lemma 4.6 Let k , m , and n be integers such that k is a divisor of the product mn . If k is relatively prime to m , then k is a divisor of n . If k is prime, then it divides either m or n (or both).

Proof. Suppose k is relatively prime to m . By Proposition 4.1 there exist integers A and B such that $1 = Am + Bk$, and so $n = Amn + Bkn$. Since k is a divisor of both Amn and Bkn , it follows that k divides their sum n . ■

Corollary 4.7 Let m and n be relatively prime integers. If m and n are divisors of k , then so is mn a divisor of k .

Proof. Suppose $k = k_1 m$. Since n is a divisor of $k = k_1 m$ and n is relatively prime to m , it follows from Lemma 4.6 that n is a divisor of k_1 . If $k_1 = k_2 n$, then $k = k_1 m = k_2 nm$, and so mn is a divisor of k . ■

Let m_1, m_2, \dots, m_n be integers which are not all zero.¹ If d is a divisor of each m_i , then d is a common divisor of the set $\{m_1, m_2, \dots, m_n\}$. If, in addition, d is divisible by every common divisor of this set, then d is that set's greatest common divisor (GCD) or highest common factor (HCF).

Theorem 4.8 Let k, m_1, m_2, \dots, m_n be integers not all of which are zero. Then there exist integers x_1, x_2, \dots, x_n such that

$$\text{GCD}(m_1, m_2, \dots, m_n) = x_1 m_1 + x_2 m_2 + \dots + x_n m_n.$$

¹The rest of this section consists of optional material which will not be used until well into Chapter 13.

Proof. Let S be the set of all the integer combinations of the form

$$x_1 m_1 + x_2 m_2 + \cdots + x_n m_n$$

where each $x_i \in \mathbb{Z}$. The choice of $x_1 = x_2 = \cdots = x_n = 0$ demonstrates that, at the very least, 0 is an element of S . If not all the m_i 's are zero, then the set contains a nonzero element, say r . If r is negative, then $-r$ is a positive element of S and hence it may be assumed that r is a positive number in S .

Let S^+ denote the set of all the positive integers of S . Since S is known to be nonempty, it follows from version 5 of the Principle of Mathematical Induction (Appendix F) that S^+ has a minimum element, say s . Next we show, by contradiction, that S is the set of all the multiples of s . Suppose t is an element of S that is not a multiple of s . By the division algorithm, there exist numbers q and r such that $0 \leq r < s$ and $t = qs + r$. But then $r = t - qs \in S$, contradicting the fact that $r \neq s$ since $0 \leq r < s$. Thus, every element of S is a multiple of s . Since s is a linear combination of the m_i 's, so is every multiple of s . Thus, the set of multiples of s is equal to the set of integer combinations of the m_i 's.

Since the elements of S are, by their definition, integer combinations of the m_i 's, it follows that s is the smallest positive integer that can be expressed as an integer combination of the m_i 's: say $s = x_1 m_1 + x_2 m_2 + \cdots + x_n m_n$. Since g divides each m_i , it follows that either

$$g \mid \sum_{i=1}^n x_i m_i$$

or $g \nmid s$. Because s is a minimum integer combination of m_1, m_2, \dots, m_n , it follows that $s \leq g$. Hence, since s is positive,

$$g = s = x_1 m_1 + x_2 m_2 + \cdots + x_n m_n. \quad \blacksquare$$

Exercises 4.2

Find the greatest common divisor of the pairs of integers in Exercises 4.2.1 to 4.2.6.

- | | |
|----------------|------------------------------|
| 1. 0 and 365 | 4. 3,367 and 4,277 |
| 2. 1 and 3,600 | 5. 123,456 and 862,091 |
| 3. 36 and 48 | 6. 14,540,165 and 85,050,243 |

Find the multiplicative inverses of the elements of \mathbb{Z}_{67} in Exercises 4.2.7 to 4.2.11.

7. 25 8. 66 9. 41 10. 37 11. 2

Find the multiplicative inverses of the elements of \mathbb{Z}_{73} in Exercises 4.2.12 to 4.2.16.

12. 25 13. 72 14. 41 15. 33 16. 2

17. Find the multiplicative inverse of 4,096 in $\mathbb{Z}_{65,537}$.
18. Find the multiplicative inverse of 1,000 in $\mathbb{Z}_{65,537}$.
19. Find the multiplicative inverse of each of the invertible elements of \mathbb{Z}_{12} .
20. Find the multiplicative inverse of each of the invertible elements of \mathbb{Z}_{18} .
21. Does the equation $399x + 703y = 114$ have an integer solution in x and y ? If so, find one; if not explain why not. (Hint: use Proposition 4.1.)
22. Does the equation $399x + 703y = 115$ have an integer solution in x and y ? If so, find one; if not, explain why not. (Hint: use Proposition 4.1.)
23. Suppose m and n are integers; characterize those integers k for which the equation $mx + ny = k$ has integer solutions in x and y . Prove your answer.
24. If $g = (m, n)$, where m and n are integers, show that g is the smallest positive integer that can be expressed in the form $Am + Bn$ where A and B vary over all integers.
25. Explain why it suffices to prove Proposition 4.1 for positive integers.
26. Prove that m has a multiplicative inverse in \mathbb{Z}_n if and only if it is relatively prime to n .
27. Prove that if m and n are two integers, then (m, n) is divisible by every common divisor of m and n .
28. Let m and n be relatively prime positive integers. Prove that there exists an integer k_0 such that for any integer $k > k_0$, $k = Am + Bn$ for some positive integers A and B .
29. Prove that if p is a prime and a and b are any integers such that $a^2 \equiv b^2 \pmod{p}$, then $a \equiv \pm b \pmod{p}$. Is this also true when p is not a prime?
30. Prove that $1 \cdot 2 \cdot 3 \cdots (p-1) \equiv -1 \pmod{p}$ whenever p is a prime.

31. The smallest positive multiple of both the integers k and m is called their least common multiple.
- (a) Prove that every common multiple of two integers is divisible by their least common multiple.
- (b) Prove that the least common multiple of any two positive integers k and m is $km/(k, m)$.
32. Let m_1, m_2, \dots, m_r denote r positive integers that are pairwise relatively prime, and let a_1, a_2, \dots, a_r denote any r integers. Then the congruences $x \equiv a_i \pmod{m_i}$, $i = 1, 2, \dots, r$, have common solutions. Any two solutions are congruent modulo $m_1 m_2 \cdots m_r$. (This is known as the *Chinese Remainder Theorem*.)
33. Suppose a/b is a rational zero of the equation

$$a_0 x^n + a_1 x^{n-1} + \cdots + a_{n-1} x + a_n = 0,$$

where a and b are relatively prime integers and a_0, a_1, \dots, a_n are arbitrary integers. Prove that a is a divisor of a_n and that b is a divisor of a_0 .

34. Prove that the equation $3x^3 - 3x^2 + 17x - 4 = 0$ has no rational roots.
35. Let a and n be any positive integers such that a is not an n -th power of any integer. Prove that $\sqrt[n]{a}$ is not a rational number.
36. Let a and b be nonzero integers and $g = (a, b)$. Prove that $(a/g, b/g) = 1$.
37. Let a , b , and c be any integers such that $g = (a, b)$ is a divisor of c . Prove that if x_0, y_0 are any integers such that $ax_0 + by_0 = (a, b)$, then the complete solution set of the equation $ax + by = c$ is

$$\left\{ (x, y) \mid x = (c/g)x_0 + (b/g)t, y = (c/g)y_0 - (a/g)t, t \in \mathbb{Z} \right\}.$$

38. Prove that if p is a prime and $\alpha, \beta \in \sqrt[p]{1}$ and $\alpha \neq 1$, then there exists an integer m such that $\alpha^m = \beta$.

4.3 Radicals in Modular Arithmetic

It might be of interest to consider briefly the issue of radicals in the context of modular arithmetic. In analogy with the more conventional number systems, if a is in \mathbb{Z}_n , then \sqrt{a} is the set of all the elements x of \mathbb{Z}_n such that $x^2 \equiv a \pmod{n}$. Thus, in \mathbb{Z}_5 , $\sqrt{0} = \{0\}$, $\sqrt{1} = \{1, 4\}$ and $\sqrt{4} = \{2, 3\}$, whereas 2 and 3 have no square roots in \mathbb{Z}_5 . Similarly, in \mathbb{Z}_7 $\sqrt{1} = \{1, 6\}$, $\sqrt{2} = \{3, 4\}$, and $\sqrt{4} = \{2, 5\}$, whereas 3, 5, and 6 have no square roots in \mathbb{Z}_7 . It is interesting to note that in \mathbb{Z}_5

$$\sqrt{-1} \equiv \sqrt{4} = \{2, 3\} \pmod{5}$$

whereas $\sqrt{-1} \equiv \sqrt{6}$ does not exist modulo 7.

The answer to the question of just which elements of \mathbb{Z}_p , p prime, do possess square roots in \mathbb{Z}_p is the subject of the *Law of Quadratic Reciprocity*. This theorem, conjectured by both Euler and Legendre and first proved by Gauss, is one of the most important theorems of number theory. It is discussed in detail in Chapter 12.

Higher order radicals are defined in a manner similar to that in which the square roots were defined. The question of existence here is much less understood, though.

Exercises 4.3

Determine which of the elements of \mathbb{Z}_n have square roots in \mathbb{Z}_n for the values of n in Exercises 4.3.1 to 4.3.4.

1. 7 2. 10 3. 13 4. 17

Determine which of the elements of \mathbb{Z}_n have cube roots in \mathbb{Z}_n for the values of n in Exercises 4.3.5 to 4.3.8.

5. 7 6. 11 7. 13 8. 17

9. Let p be any prime integer, and let a and b be nonzero elements of \mathbb{Z}_p . Show that ab has a square root in \mathbb{Z}_p if and only if either both a and b have such square roots or both a and b fail to have such square roots.

4.4 The Fundamental Theorem of Arithmetic

The Euclidean algorithm of Section 4.2 provides us with the means for proving the Fundamental Theorem of Arithmetic which states that every positive integer can be factored into the product of prime numbers in an essentially unique way.

Theorem 4.9 Let n be a positive integer that is greater than 1. Then there exist prime numbers $p_1 < p_2 < \cdots < p_h$ and positive integers r_1, r_2, \dots, r_h such that

$$n = p_1^{r_1} p_2^{r_2} \cdots p_h^{r_h}.$$

Moreover, if $q_1 < q_2 < \cdots < q_k$ is another list of primes and s_1, s_2, \dots, s_k is another list of positive integers such that

$$n = q_1^{s_1} q_2^{s_2} \cdots q_k^{s_k},$$

then $h = k$, $p_i = q_i$, and $r_i = s_i$ for $i = 1, 2, \dots, h$.

Proof. We first prove the existence of such a factorization into primes by induction on n . If $n = 2$, we can clearly use $h = 1$, and $p_1 = 2$. Let n be any positive integer such that every smaller integer that exceeds 1 has a factorization into primes. If n is a prime, then we can again set $h = 1$ and $p_1 = n$ to get a trivial factorization of n into primes. If n is not prime, say $n = n_1 n_2$ for some positive integers n_1 and n_2 both bigger than 1, then, by the induction hypothesis, n_1 and n_2 both have prime factorizations, and the product of these two factorizations yields a factorization of n into primes.

The uniqueness of the prime factorizations is also demonstrated by induction. If n is prime, it cannot be expressed as the product of other primes, and so $n = n$ is the only prime factorization of n . Let n be a positive integer such that every smaller integer that exceeds 1 has a unique prime factorization. Suppose now that

$$n = p_1^{r_1} p_2^{r_2} \cdots p_h^{r_h} = q_1^{s_1} q_2^{s_2} \cdots q_k^{s_k}.$$

A repeated application of Lemma 4.6 leads us to the conclusion that $p_i = q_i \geq q_1$ for some $i = 1, 2, \dots, k$. A symmetrical argument allows us to conclude that, in fact, $p_1 = q_1$. Consequently,

$$n/p_1 = p_1^{r_1-1} p_2^{r_2} \cdots p_h^{r_h} = q_1^{s_1-1} q_2^{s_2} \cdots q_k^{s_k}.$$

The theorem now follows from an application of the induction hypothesis (of the uniqueness of factorization) to the smaller number n/p_1 . ■

Exercises 4.4

Find the prime factorization of the numbers in Exercises 4.4.1 to 4.4.5.

- | | | |
|-----------------|-----------------|--------------|
| 1. 1,000,000 | 3. $2^{25} + 1$ | 5. 1,048,576 |
| 2. $2^{10} + 1$ | 4. 53,357 | |

6. Show that the sum, difference, and product of any two elements of the set

$$\mathbb{Z}[\sqrt{-5}] = \{ a + b\sqrt{-5} \mid a \text{ and } b \text{ are real integers} \}$$

is also in that set.

7. For any element $z = a + b\sqrt{-5}$ of the set $\mathbb{Z}[\sqrt{-5}]$ above, define $N(z) = a^2 + 5b^2$.
- Prove that $N(z) = 1$ if and only if $z = \pm 1$.
 - Prove that $N(z) \neq 3$ for all $z \in \mathbb{Z}[\sqrt{-5}]$.
 - Find all the solutions of $N(z) = 9$ in $\mathbb{Z}[\sqrt{-5}]$.
 - Prove that $N(zw) = N(z)N(w)$.
8. A nonzero element p of the set $\mathbb{Z}[\sqrt{-5}]$ is said to be *prime* if it is not ± 1 , and if whenever $p = zw$ for some $z, w \in \mathbb{Z}[\sqrt{-5}]$ we may conclude that either $z = \pm 1$ or $w = \pm 1$. Show that
- 3 , $2 + \sqrt{-5}$, and $2 - \sqrt{-5}$ are all prime elements of $\mathbb{Z}[\sqrt{-5}]$, and
 - 9 can be factored into primes of $\mathbb{Z}[\sqrt{-5}]$ in two different ways.
9. Prove that there is an infinite number of prime integers.
10. Prove that every element of $\mathbb{Z}[\sqrt{-5}]$ is expressible as the product of a finite number of primes.
11. Prove that there is an infinite number of primes in $\mathbb{Z}[\sqrt{-5}]$.
12. A positive integer is said to be a *perfect number* if it equals the sum of its proper divisors. For example, 6 and 28 are perfect because $6 = 1 + 2 + 3$ and $28 = 1 + 2 + 4 + 7 + 14$. Prove that if n is an integer such that $2^n - 1$ is a prime integer, then $(2^n - 1)2^{n-1}$ is a perfect number. Use this to find at least two more perfect numbers. (As of the writing of this text only 48 such perfect numbers have been found, the largest having $n = 57,885,161$.)

13. A positive integer is said to be *multiplicatively perfect* if it equals the product of all of its proper divisors. For example, 6 and 10 are multiplicatively perfect since $6 = 1 \cdot 2 \cdot 3$ and $10 = 1 \cdot 2 \cdot 5$. Find a simple characterization of all the multiplicatively perfect integers.

Find the number of distinct positive divisors of the numbers in Exercises 4.4.14 to 4.4.16.

14. $3^5 5^4$

15. 12^{12}

16. $p_1^{r_1} p_2^{r_2} \cdots p_h^{r_h}$ where p_1, p_2, \dots, p_h are distinct primes

17. Express the greatest common divisor and least common multiple of any two integers in terms of their factorizations into prime powers, and then redo Exercise 4.2.31.

18. Find an integer p such that p is a prime in \mathbb{Z} but $p + 0 \cdot \sqrt{-5}$ is not prime in $\mathbb{Z}[\sqrt{-5}]$.

Chapter Summary

The modular number systems \mathbb{Z}_p (p prime) share many of the properties of the rational, real, and complex number systems. They are closed with respect to the four arithmetic operations, and their elements may or may not possess square roots, cube roots, etc. In particular the question of the existence of multiplicative inverses in modular arithmetic is resolved by an application of the well-known concept of the greatest common divisor of integers. As long as this tool was under discussion we went ahead and applied it to prove the unique factorization of integers.

Chapter Review Exercises

Mark the following statements true or false.

1. $2^8 \equiv 8^2 \pmod{2^6}$.
2. 3 is a root of the congruence $x^5 + 2x + 1 \equiv 0 \pmod{50}$.
3. There exist integers x and y such that $25x + 137y = 1$.
4. 62 has a multiplicative inverse in \mathbb{Z}_{91} .
5. The multiplicative inverse of 63 in \mathbb{Z}_{100} is 25.
6. The prime factorization of 2,730 is $2 \cdot 3 \cdot 5 \cdot 91$.
7. $(-12, -18) = 6$.

New Terms

Chinese Remainder Theorem, 69	Law of Quadratic Reciprocity, 70
common divisor, 62	modular arithmetic, 57
congruent modulo n , 57	multiplicatively perfect, 73
Euclidean algorithm, 63	perfect number, 72
greatest common divisor, 62	relatively prime, 63
highest common factor, 62	

Supplementary Exercises

1. Write a computer script that will find the greatest common divisor of any two integers.
2. Write a computer script that will find the multiplicative inverse of any element in \mathbb{Z}_p for any prime p .
3. Write a computer script that will solve any equation $f(x) = 0$ in \mathbb{Z}_p for any prime p .
4. Write a computer script that will solve any pair of simultaneous linear equations in two unknowns in \mathbb{Z}_p for any prime p .
5. Write a computer script that will factor any integer into primes.
6. Write a computer script that will list all the primes up to some given integer n .
7. Write a computer script that will list all the primes in $\mathbb{Z}[\sqrt{-5}]$.
8. If n is any integer, let $\mathbb{Z}[\sqrt{n}] = \{a + b\sqrt{n} \mid a, b \in \mathbb{Z}\}$. Investigate the question of unique factorization in $\mathbb{Z}[\sqrt{n}]$ for various specific values of n .
9. Compute the number of digits in the largest known perfect number

$$2^{57,885,161}(2^{57,885,161} - 1).$$

Chapter 5



THE BINOMIAL THEOREM AND MODULAR POWERS

THE WELL-KNOWN Binomial Theorem is proved in this chapter. In addition to its intrinsic interest, this theorem also leads to a simple proof of Fermat's Theorem which, in turn, is useful for evaluating powers in arithmetic modulo a prime.

5.1 The Binomial Theorem

The formula $(a + b)^2 = a^2 + 2ab + b^2$ is, of course, standard fare in high school algebra. We are concerned here with its generalization to higher exponents. The search for this generalization begins with the successive expansions of the third and fourth powers of the binomial $(a + b)$. These are easily obtained recursively as follows:

$$\begin{aligned}(a + b)^3 &= (a + b)^2(a + b) = (a^2 + 2ab + b^2)(a + b) \\&= a^3 + 2a^2b + ab^2 + a^2b + 2ab^2 + b^3 \\&= a^3 + (2 + 1)(a^2b) + (1 + 2)ab^2 + b^3 \\&= a^3 + 3a^2b + 3ab^2 + b^3\end{aligned}$$

and

$$\begin{aligned}(a + b)^4 &= (a + b)^3(a + b) = (a^3 + 3a^2b + 3ab^2 + b^3)(a + b) \\&= a^4 + 3a^3b + 3a^2b^2 + ab^3 + a^3b + 3a^2b^2 + 3ab^3 + b^4 \\&= a^4 + (3 + 1)(a^3b) + (3 + 3)a^2b^2 + (1 + 3)b^3 + b^4 \\&= a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4.\end{aligned}$$

These low-order examples suggest that the expansion of $(a + b)^n$ consists of the sum of terms of the form $c_{n,k} a^{n-k} b^k$ where $n \geq k \geq 0$, and where the coefficient $c_{n,k}$ is a positive integer. Moreover, $c_{n,0} = c_{n,n} = 1$, and for each other k , $c_{n,k}$ is the sum of two coefficients of the expansion of $(a + b)^{n-1}$. More precisely, $c_{n,k} = c_{n-1,k} + c_{n-1,k-1}$.

The above observations motivate the following inductive definition as well as the subsequent theorem. For any integers $n \geq k \geq 0$ the *binomial coefficient* $\binom{n}{k}$, which is the traditional way of writing $c_{n,k}$, is defined by

$$\binom{n}{0} = \binom{n}{n} = 1 \quad (5.1)$$

and

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1} \quad \text{for } n > k > 0. \quad (5.2)$$

Thus, by Equation 5.1,

$$\binom{0}{0} = \binom{1}{0} = \binom{1}{1} = \binom{2}{0} = \binom{2}{2} = 1,$$

whereas by Equation 5.2,

$$\binom{2}{1} = \binom{1}{1} + \binom{1}{0} = 1 + 1 = 2.$$

Similarly,

$$\begin{aligned} \binom{3}{0} &= \binom{3}{3} = \binom{4}{0} = \binom{4}{4} = 1, & \binom{3}{1} &= \binom{2}{1} + \binom{2}{0} = 2 + 1 = 3, \\ \binom{3}{2} &= \binom{2}{2} + \binom{2}{1} = 1 + 2 = 3, & \binom{4}{1} &= \binom{3}{1} + \binom{3}{0} = 3 + 1 = 4, \\ \binom{4}{2} &= \binom{3}{2} + \binom{3}{1} = 3 + 3 = 6, & \binom{4}{3} &= \binom{3}{3} + \binom{3}{2} = 1 + 3 = 4. \end{aligned}$$

Since $(a + b)^0 = 1$ and $(a + b)^1 = (a + b)$, these numbers are now easily seen to agree with the coefficients of the expansions of $(a + b)^n$ for $n = 0, 1, 2, 3, 4$.

Theorem 5.3 (The Binomial Theorem) Let n be a nonnegative integer. Then

$$(a + b)^n = \binom{n}{0}a^n + \binom{n}{1}a^{n-1}b + \cdots + \binom{n}{k}a^{n-k}b^k + \cdots + \binom{n}{n}b^n.$$

Proof. We proceed by induction on n , the cases $n = 0, 1, 2, 3, 4$ having been verified above. Assuming the theorem for $n = m$, we expand

$$\begin{aligned} (a + b)^{m+1} &= (a + b)^m(a + b) \\ &= \left[\binom{m}{0}a^m + \binom{m}{1}a^{m-1}b + \cdots + \binom{m}{k}a^{m-k}b^k + \cdots + \binom{m}{m}b^m \right] (a + b) \\ &= \binom{m}{0}a^{m+1} + \binom{m}{1}a^m b + \binom{m}{2}a^{m-1}b^2 + \cdots + \binom{m}{k}a^{m+1-k}b^k + \cdots \\ &\quad + \binom{m}{m}ab^m + \binom{m}{0}a^m b + \binom{m}{1}a^{m-1}b^2 + \cdots \\ &\quad + \binom{m}{k-1}a^{m-(k-1)}b^k + \cdots + \binom{m}{m}b^{m+1} \\ &= \binom{m}{0}a^{m+1} + \left[\binom{m}{1} + \binom{m}{0} \right] a^m b + \left[\binom{m}{2} + \binom{m}{1} \right] a^{m-1}b^2 + \cdots \\ &\quad + \left[\binom{m}{k} + \binom{m}{k-1} \right] a^{m+1-k}b^k + \cdots + \binom{m}{m}b^{m+1} \\ &= \binom{m+1}{0}a^{m+1} + \binom{m+1}{1}a^m b + \binom{m+1}{2}a^{m-1}b^2 + \cdots \\ &\quad + \binom{m+1}{k}a^{m+1-k}b^k + \cdots + \binom{m+1}{m+1}b^{m+1}. \end{aligned}$$

This completes the induction step and the proof. ■

Equation 5.2 is known as *Pascal's Identity*. It was visualized by Pascal in the form of a triangular array (*Pascal's Triangle*, Figure 5.1) wherein each number is the sum of the two entries directly above it and to its left (whenever these entries exist). The entries on the $(n + 1)$ -th diagonal line (from bottom left to top right) are the coefficients that appear in the expansion of $(a + b)^n$. These triangular arrays did not originate with Pascal. Figure 5.2 displays a reproduction of such an array that appeared over three centuries before his birth.

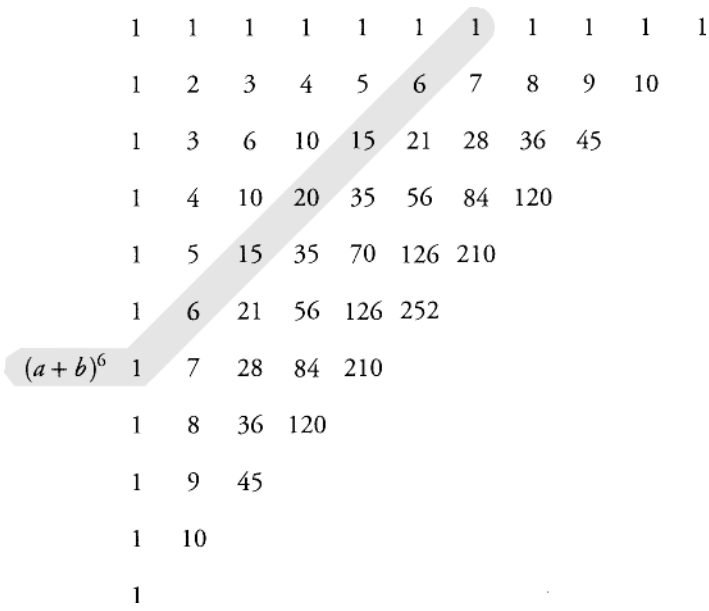


Figure 5.1 Pascal's Triangle



Figure 5.2 Pascal's Triangle as depicted in a Chinese book in 1303

It should be noted that the Binomial Theorem is valid for all the number systems we have studied so far. It is immaterial whether a and b are to be interpreted as real, complex, or modular numbers. Thus,

$$(2 + i)^3 = 2^3 + 3 \cdot 2^2 \cdot i + 3 \cdot 2 \cdot i^2 + i^3 = 8 + 12i - 6 - i = 2 + 11i.$$

Similarly, in \mathbb{Z}_3 ,

$$(a + 2)^3 = a^3 + 3 \cdot a^2 \cdot 2 + 3 \cdot a \cdot 2^2 + 2^3 \equiv a^3 + 0 + 0 + 2 = a^3 + 2.$$

In \mathbb{Z}_4 ,

$$(a + b)^4 = a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4 \equiv a^4 + 2a^2b^2 + b^4.$$

For the purposes of the remainder of this section it may be assumed that the a and the b of the Binomial Theorem are real numbers.

While Pascal's Identity provides an effective procedure for computing the binomial coefficients, it would be clearly handy to have a more direct method. To obtain such a formula, the a of the Binomial Theorem is replaced by 1, and, since we are about to differentiate, the b is replaced by x . The Binomial Theorem then assumes the form

$$(1 + x)^n = 1 + \binom{n}{1}x + \binom{n}{2}x^2 + \cdots + \binom{n}{k}x^k + \cdots + \binom{n}{n}x^n. \quad (5.4)$$

When Equation 5.4 is differentiated k times with respect to x , we obtain

$$\begin{aligned} n(n-1)(n-2) \cdots (n-k+1)(1+x)^{n-k} \\ = k(k-1) \cdots 2 \cdot 1 \binom{n}{k} + d_1x + d_2x^2 + \cdots + d_{n-k}x^{n-k} \end{aligned} \quad (5.5)$$

where d_1, d_2, \dots, d_{n-k} are some integers whose exact values turn out to be immaterial. The substitution of $x = 0$ in Equation 5.5 yields

$$n(n-1)(n-2) \cdots (n-k+1) = k(k-1) \cdots 2 \cdot 1 \binom{n}{k},$$

or

$$\binom{n}{k} = \frac{n(n-1)(n-2) \cdots (n-k+1)}{k(k-1) \cdots 2 \cdot 1}. \quad (5.6)$$

Accordingly, for $n = 6$,

$$\begin{aligned} \binom{6}{1} &= \frac{6}{1} = 6, & \binom{6}{2} &= \frac{6 \cdot 5}{2 \cdot 1} = 15, & \binom{6}{3} &= \frac{6 \cdot 5 \cdot 4}{3 \cdot 2 \cdot 1} = 20, \\ \binom{6}{4} &= \frac{6 \cdot 5 \cdot 4 \cdot 3}{4 \cdot 3 \cdot 2 \cdot 1} = 15, & \binom{6}{5} &= \frac{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 6, & \binom{6}{6} &= \frac{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 1, \end{aligned}$$

so that

$$(a + b)^6 = a^6 + 6a^5b + 15a^4b^2 + 20a^3b^3 + 15a^2b^4 + 6a^5b + b^6.$$

While the formula for $\binom{n}{k}$ given in Equation 5.6 is convenient for numerical computations, it is quite frequently better to use the equivalent expression

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \quad (5.7)$$

where $k! = k(k-1)(k-2) \cdots 2 \cdot 1$ for any positive integer k and $0! = 1$ (Exercise 5.1.15).

The binomial coefficients $\binom{n}{k}$ are the subject of many surprising identities and we shall describe several methods for proving them. The most elementary approach uses Equation 5.6. Such is the case in the proof of the identity

$$\binom{n}{k+1} = \frac{n-k}{k+1} \binom{n}{k}. \quad (5.8)$$

For

$$\begin{aligned} \binom{n}{k+1} &= \frac{n(n-1) \cdots [n-(k+1)+1](k+1)}{k(k-1) \cdots 2 \cdot 1} = \frac{n(n-1) \cdots (n-k)}{(k+1)k(k-1) \cdots 2 \cdot 1} \\ &= \frac{n-k}{k+1} \cdot \frac{n(n-1)(n-2) \cdots (n-k+1)}{k(k-1) \cdots 2 \cdot 1} = \frac{n-k}{k+1} \binom{n}{k}. \end{aligned}$$

Equation 5.8 was put to very good use by Newton in his groundbreaking work on the extension of the Binomial Theorem to negative and fractional values of n .

As demonstrated by the proof of the Binomial Theorem, Equation 5.2 (Pascal's Identity) is very useful as it lays the groundwork for many inductive proofs of other identities. However, many of these identities are subject to shorter proofs by another method,

commonly called the *method of generating functions*. Consider, for example, the identity

$$2^n = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{k} + \cdots + \binom{n}{n-1} + \binom{n}{n} \quad (5.9)$$

for $n \geq 0$. While this identity can be proven directly, though somewhat laboriously, by mathematical induction, it can also be verified by substituting $a = b = 1$ in the Binomial Theorem. Thus, this method calls for the recognition of some functional identity which, upon the replacement of the variable(s) by some cleverly chosen values, yields the required identity. In fact, the above derivation of Equation 5.6 is also an instance of the method of generating functions, as is the following. If both sides of Equation 5.4 are differentiated with respect to x , we obtain

$$n(1+x)^{n-1} = \binom{n}{1} + 2\binom{n}{2}x + \cdots + k\binom{n}{k}x^{k-1} + \cdots + n\binom{n}{n}x^{n-1}.$$

The substitution of $x = 1$ now yields the identity

$$n2^{n-1} = \binom{n}{1} + 2\binom{n}{2} + \cdots + k\binom{n}{k} + \cdots + n\binom{n}{n}.$$

Finally, there is another interpretation of the binomial coefficients that allows for a completely different approach to the whole topic. For this we need to reexamine the Binomial Theorem from another point of view. Each summand in the expansion of

$$(a_1 + a_2)(b_1 + b_2)(c_1 + c_2)(d_1 + d_2) \cdots$$

has the form $a_i b_j c_k d_l \cdots$ where each of the indices i, j, k , and l assumes the values 1 or 2. When this observation is applied to the expression

$$(1+x)^n = (1+x)(1+x)(1+x)(1+x) \cdots$$

it is clear that each of the summands of this expansion has the form $X_1 X_2 \cdots X_n$ where each X_i is either 1 or x . In particular, the general summand $X_1 X_2 \cdots X_n$ is of degree 2 in x if it has the form

$$1 \cdot 1 \cdots 1 \cdot x \cdot 1 \cdots 1 \cdot x \cdot 1 \cdots 1.$$

The coefficient $\binom{n}{2}$ clearly equals the number of such summands that appear in the expansion of $(1+x)^n$. Since the number of such summands equals the number of pairs of symbols X_i, X_j that can be designated for replacement by x , it follows that we now have obtained a formula for the number of ways a pair of distinct objects can be selected from a given set of n distinct objects, namely, this number is $\binom{n}{2}$.

Similarly, $\binom{n}{3}$ equals the number of summands in the expansion of $(1+x)^n$ that have degree 3 in x , i.e., it equals the number of summands of the form

$$1 \cdot 1 \cdots 1 \cdot x \cdot 1 \cdots 1 \cdot x \cdot 1 \cdots 1 \cdot x \cdot 1 \cdots 1.$$

This number equals the number of distinct triples X_i, X_j, X_k that can be selected from X_1, X_2, \dots, X_n . In other words, $\binom{n}{3}$ equals the number of three-element sets that can be formed using the integers $1, 2, \dots, n$. This generalizes to the following statement.

Proposition 5.10 For each pair of nonnegative integers $k \leq n$, the binomial coefficient $\binom{n}{k}$ equals the number of k -element subsets that can be formed from the integers $1, 2, \dots, n$.

For example, the number of three-person committees that can be selected from a group of 25 people is

$$\binom{25}{3} = \frac{25 \cdot 24 \cdot 23}{3 \cdot 2 \cdot 1} = 2,300.$$

This point of view provides us with a new tool for proving some facts about the binomial coefficients. Let us again consider Identity 5.9. Its right-hand side now has the obvious interpretation of denoting the number of all the subsets of $\{1, 2, \dots, n\}$, classified by their cardinality. This number, however, also happens to be 2^n as can be seen by the following argument. Choosing a subset of S of $\{1, 2, \dots, n\}$ is tantamount to deciding for each $k = 1, 2, \dots, n$ whether or not k belongs to S . Since n such decisions are to be made, and since each such decision can be settled in one of two ways (to belong or not to belong), it follows that there are 2^n such subsets S .

Exercises 5.1

Expand the binomials in Exercises 5.1.1 to 5.1.6.

1. $(2x + 3y^2)^7$
2. $(3x^2 - yz^3)^5$
3. $(3x^2 - yz^3)^5$ in \mathbb{Z}_5
4. $(3x^2 - yz^3)^5$ in \mathbb{Z}_6
5. $(z^2 + 2)^6$ in \mathbb{Z}_4
6. $(1 - 2/x)^5$

7. Find the term containing a^{26} in the expansion of $(a - 4b^2c^3)^{30}$.
8. Find the coefficient of x^{18} in $(x^2 + 3/x)^{15}$.
9. Find the coefficient of x^{18} in $(2x^4 - 3x)^9$.
10. Find the coefficients of x^{-4} and x^{-5} in $(x^3 - 2/x^2)^{10}$.
11. Prove that for $m = 4n, 4n - 3, 4n - 6, \dots, -2n$, the coefficient of x^m in $(x^2 + 1/x)^{2n}$ is

$$\frac{(2n)!}{\left(\frac{4n-m}{3}\right)! \left(\frac{2n+m}{3}\right)!}.$$

12. Prove that for $1 \leq k \leq n$, $\binom{n}{k-1} \geq \binom{n}{k}$ if $k \geq (n+1)/2$ and $\binom{n}{k-1} \leq \binom{n}{k}$ if $k \leq (n+1)/2$.
13. Show that the middle coefficient(s) of the expansion of $(1+x)^n$ is (are) the largest.
14. Which power of x has the largest coefficient in the expansion of $(2+3x)^{17}$?

Prove the identities in Exercises 5.1.15 to 5.1.19 for any two integers k and n such that $0 \leq k \leq n$.

15. $\binom{n}{k} = (n!)/(k!(n-k)!)$ 16. $\binom{n}{k} = \binom{n}{n-k}$
17. $\binom{n}{r}\binom{r}{k} = \binom{n}{k}\binom{n-k}{r-k} \quad (n \geq r \geq k)$
18. $n\binom{n}{r} = (r+1)\binom{n}{r+1} + r\binom{n}{r} = r\binom{n+1}{r+1} + \binom{n}{r+1}$
19. $\binom{n}{2}\binom{n}{r} = \binom{r+2}{2}\binom{n}{r+2} + 2\binom{r+1}{2}\binom{n}{r+1} + \binom{r}{2}\binom{n}{r} = \binom{r}{2}\binom{n+2}{r+2} + 2r\binom{n+1}{r+2} + \binom{n}{r+2}$
20. Prove that the coefficient of the middle term of $(1+x)^{2n}$ equals the sum of the coefficients of the two middle terms of $(1+x)^{2n-1}$.
21. Prove that $2n < \binom{2n}{n} < 4^n$ for $n > 1$.
22. Prove that the product of any k consecutive positive integers is an integer multiple of $k!$.
23. Prove that the number of different 0-1 strings that may be formed with p 0's and q 1's in which no two 1's are consecutive is $\binom{p+q}{q}$. For example, if $p = q = 3$, these strings are 010101, 100101, 101001, and 101010.
24. Prove that $\binom{2n}{n}$ is even for $n > 0$.

Prove the identities in Exercises 5.1.25 to 5.1.36.

25. $\binom{n}{0} - \binom{n}{1} + \binom{n}{2} - \dots + (-1)^n \binom{n}{n} = 0 \quad (n > 0)$
26. $\binom{n}{1} + \binom{n}{3} + \binom{n}{5} + \dots = 2^{n-1} \quad (n > 0)$

27. $\binom{n}{1} - 2\binom{n}{2} + 3\binom{n}{3} - \cdots + (-1)^{n-1} n\binom{n}{n} = 0 \quad (n > 1)$
28. $\binom{n}{0} + \frac{1}{2}\binom{n}{1} + \frac{1}{3}\binom{n}{2} + \cdots + \frac{1}{n+1}\binom{n}{n} = \frac{2^{n+1}-1}{n+1}$
29. $\binom{n}{0} - \frac{1}{2}\binom{n}{1} + \frac{1}{3}\binom{n}{2} - \cdots + \frac{(-1)^n}{n+1}\binom{n}{n} = \frac{1}{n+1}$
30. $\binom{n}{0}^2 + \binom{n}{1}^2 + \binom{n}{2}^2 + \cdots + \binom{n}{n}^2 = \binom{2n}{n}$
31. $\binom{n}{0} - \binom{n}{2} + \binom{n}{4} - \binom{n}{6} + \cdots = 2^{n/2} \cos(n\pi/4)$
32. $\binom{n}{0} + \binom{n+1}{1} + \binom{n+2}{2} + \cdots + \binom{n+k}{k} = \binom{n+k+1}{k}$
33. $\binom{n}{0}\binom{m}{k} + \binom{n}{1}\binom{m}{k-1} + \binom{n}{2}\binom{m}{k-2} + \cdots + \binom{n}{k}\binom{m}{0} = \binom{n+m}{k}$
34. $\binom{n}{0}\binom{m}{0} + \binom{n}{1}\binom{m}{1} + \binom{n}{2}\binom{m}{2} + \cdots + \binom{n}{m}\binom{m}{m} = \binom{n+m}{m}$
35. $\binom{n}{1}/\binom{n}{0} + 2\binom{n}{2}/\binom{n}{1} + 3\binom{n}{3}/\binom{n}{2} + \cdots + n\binom{n}{n}/\binom{n}{n-1} = \binom{n+1}{2}$
36. $[(\binom{n}{0} + \binom{n}{1})][(\binom{n}{1} + \binom{n}{2})] \cdots [(\binom{n}{n-1} + \binom{n}{n})] = \binom{n}{1}\binom{n}{2} \cdots \binom{n}{n} \frac{(n+1)^n}{n!}$

Simplify the expressions in Exercises 5.1.37 to 5.1.39.

37. $\binom{n}{1} + 2^2\binom{n}{2} + 3^2\binom{n}{3} + \cdots + n^2\binom{n}{n}$
38. $\binom{n}{1} + 2^3\binom{n}{2} + 3^3\binom{n}{3} + \cdots + n^3\binom{n}{n}$
39. $\binom{n}{0}\binom{n}{1} + \binom{n}{1}\binom{n}{2} + \binom{n}{2}\binom{n}{3} + \cdots + \binom{n}{n-1}\binom{n}{n}$
40. The *Fibonacci numbers* F_n are defined inductively as $F_1 = F_2 = 1$ and $F_{n+2} = F_{n+1} + F_n$ for $n \geq 1$. Prove that

$$F_n + 1 = 1 + \binom{n-1}{1} + \binom{n-2}{2} + \cdots + \binom{n-m}{m}$$

for $n \geq 0$, where $m = n/2$ if n is even and $m = (n-1)/2$ if n is odd.

41. Prove that if m is an odd integer, then $\binom{m-1}{2} \equiv 1 \pmod{m}$.
42. Prove that if p is a prime and m is an integer such that $m \not\equiv 0 \pmod{p}$, then $\binom{p^k m}{p^k} \not\equiv 0 \pmod{p}$ for all positive integers k .
43. Prove that if k is any positive integer and p is a prime, then $\binom{p^k-1}{j} \equiv (-1)^j \pmod{p}$ for $j = 0, 1, 2, \dots, p^k - 1$.
44. The grid $G_{m,n}$ consists of the graphs of the lines $x = i$ and $y = j$ for $i = 0, 1, \dots, m$ and $j = 0, 1, \dots, n$ with $0 < x < m$ and $0 < y < n$ (see Figure 5.3). Prove that $G_{m,n}$ contains $\binom{m+1}{2}\binom{n+1}{2}$ rectangles.
45. A path from $(0, 0)$ to (m, n) in the grid $G_{m,n}$ consists of a sequence of integer points $(x_0, y_0), (x_1, y_1), (x_2, y_2), \dots, (x_{m+n}, y_{m+n})$ such that $(x_0, y_0) = (0, 0)$, $(x_{m+n}, y_{m+n}) = (m, n)$, and, for $k = 1, 2, \dots, m+n$, either $(x_k, y_k) = (x_{k-1} + 1, y_{k-1})$ or $(x_k, y_k) =$

$(x_{k-1}, y_{k-1} + 1)$. Prove that the number of distinct paths from $(0, 0)$ to (m, n) paths in $G_{m,n}$ is $\binom{m+n}{n}$.

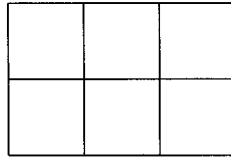


Figure 5.3 The grid $G_{3,2}$

46. How many triangles are contained in Figure 5.4(a)?
 47. How many triangles are contained in Figure 5.4(b)?

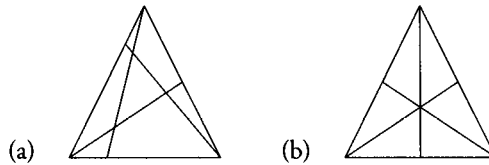


Figure 5.4 Triangle counting problems

48. Find an error in Figure 5.2.

5.2 Fermat's Theorem and Modular Exponents

We now return to modular arithmetic and address the issue of exponentiation. Consider the infinite sequence

$$2, 2^2, 2^3, 2^4, 2^5, 2^6, 2^7, 2^8, 2^9, \dots \quad (5.11)$$

in \mathbb{Z}_5 . Since \mathbb{Z}_5 contains only five elements, it is clear that the actual values of these powers must display many repetitions. In fact, when these powers are evaluated, the sequence is transformed into

$$2, 4, 3, 1, 2, 4, 3, 1, 2, \dots \quad (5.12)$$

and so Sequence 5.11, and therefore also Sequence 5.12, will cycle indefinitely through the values 2, 4, 3, and 1. In general, if $a \in \mathbb{Z}_n$, the sequence

$$a, a^2, a^3, a^4, \dots \quad (5.13)$$

is bound to eventually cycle since its individual terms run through a finite number of values, and once $a^k \equiv a^m \pmod{n}$, we must clearly have $a^{k+s} \equiv a^{m+s} \pmod{n}$ for $s = 0, 1, 2, \dots$. What is not obvious is that when n is a prime p , the cycling begins immediately and length of this cycle is either $p - 1$ or some divisor thereof. We now set out to prove these interesting facts.

As was observed in the previous section, the Binomial Theorem also holds in \mathbb{Z}_m . Under certain circumstances, this modular Binomial Theorem assumes a very simple form. If p is a prime number and $0 < k < p$, then the binomial coefficient $\binom{p}{k}$ is an integer that equals

$$\frac{p(p-1)\cdots(p-k+1)}{k(k-1)\cdots 1}.$$

Since p is a prime greater than k , p is relatively prime to the denominator of this fraction and so, as this fraction is known to cancel out to an integer, p must be a prime factor of this integer. In other words, $\binom{p}{k} \equiv 0 \pmod{p}$ if $0 < k < p$. When this observation is applied to the expansion of $(a+b)^p$, we obtain the following curious fact.

Proposition 5.14 If p is a prime integer and $a, b \in \mathbb{Z}_p$, then

$$(a+b)^p \equiv a^p + b^p \pmod{p}.$$

This proposition, in turn, has as its consequence one of the fundamental elementary and nonobvious theorems of number theory. It was first pointed out by Fermat as a tool in the search for perfect numbers (Exercise 4.4.12), but has since completely transcended that narrow context.

Theorem 5.15 (Fermat's Theorem) If p is any prime integer and a is any integer, then $a^p \equiv a \pmod{p}$.

Proof. We proceed by induction on a , the theorem being clearly true for $a = 0$. Assuming that the theorem holds for $a = k$, we note that

$$(k+1)^p \equiv k^p + 1^p \equiv k+1 \pmod{p}.$$

Hence, by induction, the theorem holds for all the nonnegative values of a . Since every negative integer is congruent to some positive integer modulo p , the theorem holds for all values of a . ■

Fermat's Theorem guarantees that Sequence 5.13 starts cycling immediately, a phenomenon that need not happen in \mathbb{Z}_n when n is composite (Exercise 5.2.27). It still

remains to demonstrate that the length of the cycle is a divisor of $p - 1$ and this will be done shortly.

An interesting and typical consequence of this theorem is the fact that if n is relatively prime to 7, then $n^6 - 1$ is divisible by 7. For, by Fermat's Theorem $n^7 \equiv n \pmod{7}$. Since n is relatively prime to 7, it follows that $n \not\equiv 0 \pmod{7}$ and so division by n yields $n^6 \equiv 1 \pmod{7}$, which is tantamount to saying that 7 divides $n^6 - 1$. It is clear that the same holds for every other prime.

Corollary 5.16 If p is any prime and $a \equiv 0 \pmod{p}$, then $a^{p-1} \equiv 1 \pmod{p}$.

Corollary 5.16 can be interpreted as saying that every nonzero element of \mathbb{Z}_p (p prime) is a $(p - 1)$ -th root of unity modulo p . Such roots of unity can exist in \mathbb{Z}_n for composite n as well. Thus, $3^4 \equiv 5^4 \equiv 7^2 \equiv 1 \pmod{16}$. We shall henceforth understand the term root of unity to include both the complex ones and the appropriate elements of \mathbb{Z}_n for all integers n . When the need arises, we shall refer to the latter as *modular roots of unity*.

For any modular root of unity r in \mathbb{Z}_n we define $o(r)$, the *modular order* of r in \mathbb{Z}_n , to be the least positive integer k such that $r^k \equiv 1 \pmod{n}$. Corollary 5.16 guarantees that every nonzero element of \mathbb{Z}_p , p prime, has a finite order which is in fact no greater than $p - 1$. Table 5.1 lists the orders of the nonzero elements of \mathbb{Z}_7 .

m	1	2	3	4	5	6
$o(m)$	1	3	6	3	6	2

Table 5.1 The elements of \mathbb{Z}_7 and their orders

The fact that all the orders of Table 5.1 are divisors of $6 = 7 - 1$ is no coincidence. Modular order enjoys the same properties as does the order of the complex roots of unity, and we restate these properties in this new context without proof. It is easily verified that the proofs of Proposition 2.16 and Corollary 2.17 carry over to this new context without any modifications whatsoever.

Proposition 5.17 If r is any modular root of unity in \mathbb{Z}_n and k is any integer, then $r^k \equiv 1 \pmod{n}$ if and only if k is a multiple of $o(r)$.

Corollary 5.18 Suppose r is a root of unity and a and b are any two integers. Then $r^a = r^b$ if and only if $o(r)$ is a divisor of $a - b$, and $1, r, r^2, \dots, r^{o(r)-1}$ are all distinct.

Returning to Equation 5.13, it follows from Proposition 5.17 that $o(a)$ is a divisor of $p - 1$. As the length of the repeating segment of Equation 5.13 equals $o(a)$, we may conclude that this length is a divisor of $p - 1$.

An element of \mathbb{Z}_p is said to be a *primitive root* modulo p if its order in \mathbb{Z}_p is $p - 1$. Thus, according to Table 5.1, 3 and 5 are the only primitive roots modulo 7. We shall later see (Theorem 7.17) that for every prime number p there exist primitive roots modulo p . The solutions of Exercises 5.2.14 and 5.2.16 indicate that this fact is far from obvious.

We now go on to prove some more facts about orders of roots of unity. If ζ is any root of unity and k is any integer, then it stands to reason that the order of ζ^k should depend on k and the order of ζ . Thus, if $o(\zeta) = 12$, then the orders of $\zeta^2, \zeta^3, \dots, \zeta^{11}$ are easily seen to be 6, 4, 3, 12, 2, 12, 3, 4, 6, 12, respectively. A little experimentation leads to the statement, if not the proof, of the following proposition.

Proposition 5.19 If ζ is a root of unity of order n , $o(\zeta^k) = n/(k, n)$ for all integers k .

Proof. Let g denote the greatest common divisor of k and n , and let n' and k' be integers such that $n = gn'$ and $k = gk'$. Since $g = (k, n)$, it follows that k' and n' are relatively prime. If m is any integer, then the following statements are all equivalent:

- $(\zeta^k)^m = 1$; ▪ gn' divides $gk'm$; ▪ n' divides m .
- n divides km ; ▪ n' divides $k'm$;

Hence $o(\zeta^k)$, the least positive integer m for which $(\zeta^k)^m = 1$, is also the least positive integer that is divisible by n' , namely n' itself. Thus $o(\zeta^k) = n' = ng = n/(k, n)$. ■

The relevance of common divisors to the orders of roots of unity is reinforced by the next proposition for which we will eventually find several useful applications.

Proposition 5.20 If r and s are roots of unity (both complex or both modular) and if the orders of r and s are relatively prime, then $o(rs) = o(r)o(s)$.

Proof. Let $R = o(r)$, $S = o(s)$, and $T = o(rs)$. Since $(rs)^{RS} = (r^R)^S (s^S)^R = 1$, it follows that T is a divisor of RS . Conversely, $1 = (rs)^{TS} = r^{TS} s^{ST} = r^{TS} \cdot 1 = r^{TS}$. It therefore follows from Proposition 5.17 that R is a divisor of TS . Since R and S are relatively prime, it follows from Lemma 4.6 that R is a divisor of T . A similar argument permits us to conclude that S is also a divisor of T . Since R and S are relatively prime, it now follows from Corollary 4.7 that RS is a divisor of T . Thus, we have shown that T and RS each divide the other and the proposition follows. ■

Thus, if $\zeta = \cos 2\pi/12 + i \sin 2\pi/12$, then the elements ζ^8 and ζ^9 of $\sqrt[12]{1}$ have orders 3 and 4, respectively, and the element $\zeta^5 = \zeta^8 \zeta^9$ has order $12 = 3 \cdot 4$. Similarly, in \mathbb{Z}_{19} , $\text{o}(18) = 2$, $\text{o}(7) = 3$, and $\text{o}(18 \cdot 7) = \text{o}(12) = 6$. Exercise 5.2.22 implies that the requirement of relative primeness in the above proposition is indeed necessary.

Exercises 5.2

1. Evaluate $x^{1,000}$ for each element x of \mathbb{Z}_7 .
2. Evaluate $x^{2,000}$ for each element x of \mathbb{Z}_{11} .

Solve the equations in Exercises 5.2.3 to 5.2.5 in \mathbb{Z}_7 .

$$3. \quad x^{7,777} + x + 5 \equiv 0 \qquad 4. \quad x^{6,777} + x + 5 \equiv 0 \qquad 5. \quad x^{5,777} + x + 5 \equiv 0$$

6. Solve the equation in Exercise 5.2.3 in \mathbb{Z}_{11} .
7. Solve the equation in Exercise 5.2.4 in \mathbb{Z}_{13} .
8. Solve the equation in Exercise 5.2.5 in \mathbb{Z}_{17} .
9. Prove that $n^7 - n$ is divisible by 42 for any integer n .
10. Prove that $n^{13} - n$ is divisible by 2,730 for every integer n .
11. Prove that $n^5 - n$ is divisible by 30 for all integers n and by 240 for all odd n .
12. Prove that $n^{561} \equiv n \pmod{561}$. Note that this example disproves the converse of Fermat's Theorem. Composite numbers with this property are called Carmichael numbers; 561 is the smallest Carmichael number, and it was only recently proved that there are infinitely many such numbers.
13. For any prime p , if $a^p \equiv b^p \pmod{p}$, show that $a^p \equiv b^p \pmod{p^2}$.
14. Find the units digit of 2^{400} .
15. Find all the primitive roots modulo 11.
16. Find all the primitive roots modulo 19.
17. For each prime $p < 20$ find a number that is a primitive root mod p .
18. Let p be a fixed prime and let $\text{o}(a)$ denote the order of a in \mathbb{Z}_p . Prove that if $a \neq 1$, then $1 + a + a^2 + \cdots + a^{\text{o}(a)-1} \equiv 0 \pmod{p}$.
19. Let p be a fixed prime and let $\text{o}(a)$ denote the order of a in \mathbb{Z}_p . Prove that

$$a \cdot a^2 \cdots a^{\text{o}(a)-1} \equiv (-1)^{\text{o}(a)-1} \pmod{p}.$$

20. Let p be any prime number. Prove that $(p-1)! \equiv -1 \pmod{p}$.
21. Let p be any prime number. Prove that if there are primitive roots mod p , then the product of all of them is equivalent to 1 or 2 mod p .
22. Prove that for any odd prime p there exist nonzero elements a and b of \mathbb{Z}_p such that $o(ab) \neq o(a)o(b)$.
23. The Fibonacci number F_n is defined recursively as $F_1 = F_2 = 1$ and $F_n = F_{n-1} + F_{n-2}$ for $n > 2$.

(a) Prove that

$$F_n = \frac{1}{\sqrt{5}} \left[\left(\frac{1+\sqrt{5}}{2} \right)^n - \left(\frac{1-\sqrt{5}}{2} \right)^n \right]$$

for all $n = 1, 2, \dots$ (this is known as Binet's Formula).

(b) Prove that if p is a prime distinct from 5, then $F_p \equiv \pm 1 \pmod{p}$.

24. Show that the equation $x^n + y^n \equiv z^n \pmod{3}$ has nonzero solutions in \mathbb{Z}_3 if and only if n is odd.
25. Show that the equation $x^n + y^n \equiv z^n \pmod{5}$ has nonzero solutions in \mathbb{Z}_5 if and only if n is odd.
26. For which positive integers n does the equation $x^n + y^n \equiv z^n \pmod{7}$ have nonzero solutions in \mathbb{Z}_7 ?
27. Find integers a and n such that $a \in \mathbb{Z}_n$ but the Sequence 5.13 does not cycle immediately.

5.3 The Multinomial Theorem

Having obtained the Binomial Theorem, which describes the expansion of $(a+b)^n$, it is natural to ask for analogous expressions for $(a+b+c)^n$, $(a+b+c+d)^n$, etc. It turns out that these expressions are harder to describe than to derive. We begin by changing the variables to x_1, x_2, \dots , and observe that for any v such variables and for any positive integer n , the expansion of $(x_1 + x_2 + \dots + x_v)^n$ consists of the sum of terms each of which has the form

$$c x_1^{k_1} x_2^{k_2} \dots x_v^{k_v}$$

where $k_1 + k_2 + \dots + k_v = n$ and c is a positive integer that depends on n, k_1, k_2, \dots, k_v .

Theorem 5.21 (The Multinomial Theorem) If n and v are any positive integers, then

$$(x_1 + x_2 + \cdots + x_v)^n = \sum_K \frac{n!}{k_1! k_2! \cdots k_v!} x_1^{k_1} x_2^{k_2} \cdots x_v^{k_v}$$

where K varies over all the v -tuples $K = (k_1, k_2, \dots, k_v)$ of nonnegative integers k_1, k_2, \dots, k_v such that $k_1 + k_2 + \cdots + k_v = n$.

Proof. Let n and v be fixed positive integers. As noted above, for each v -tuple $K = (k_1, k_2, \dots, k_v)$ of the above format there exists an integer c_K such that

$$(x_1 + x_2 + \cdots + x_v)^n = \sum_K c_K x_1^{k_1} x_2^{k_2} \cdots x_v^{k_v}. \quad (5.22)$$

Fix some such $K = (k_1, k_2, \dots, k_v)$ and differentiate both of the sides of Equation 5.22 k_i times with respect to x_i for each $i = 1, 2, \dots, v$. Since $k_1 + k_2 + \cdots + k_v = n$,

$$\frac{\partial^n}{\partial^{k_1} x_1 \partial^{k_2} x_2 \cdots \partial^{k_v} x_v} (x_1 + x_2 + \cdots + x_v)^n = n!.$$

Moreover, for any v -tuple (m_1, m_2, \dots, m_v) that is different from (k_1, k_2, \dots, k_v) and for which $m_1 + m_2 + \cdots + m_v = n$, we must have $m_i < k_i$ for some i so that

$$\frac{\partial^n}{\partial^{k_1} x_1 \partial^{k_2} x_2 \cdots \partial^{k_v} x_v} (x_1^{m_1} x_2^{m_2} \cdots x_v^{m_v}) = \begin{cases} k_1! k_2! \cdots k_v! & \text{if } m_i = k_i \text{ for all } i; \\ 0 & \text{otherwise.} \end{cases}$$

Hence $n! = c_K k_1! k_2! \cdots k_v!$ and

$$c_K = \frac{n!}{k_1! k_2! \cdots k_v!}. \quad \blacksquare$$

Accordingly, the coefficient of $x^4 y^3 z$ in the expansion of $(x + y + z)^8$ is $8! 4! 3! 1! = 280$.

5.4 The Euler φ -Function

Corollary 4.4 guarantees that whenever m is relatively prime to n , it has a multiplicative inverse in \mathbb{Z}_n , and according to Exercise 4.2.26 this is actually the complete answer. Namely, if m is not relatively prime to n , then it does not possess a multiplicative inverse in \mathbb{Z}_n . Thus 1 and 5 are the only elements of \mathbb{Z}_6 that have multiplicative inverses whereas 1, 3, 5, and 7 are the only elements of \mathbb{Z}_8 that have multiplicative inverses. This raises the interesting question of just how many elements of \mathbb{Z}_n do in general have such inverses. As the resulting formula has some bearing on the complex roots of unity and on other subsequent issues, it will be derived here. For any positive integer n let $\varphi(n)$ denote the number of positive integers not greater than n that are relatively prime to n . This is known as the *Euler φ -function*. As noted above, $\varphi(6) = 2$ and $\varphi(8) = 4$. It is clear that if p is a prime, then $\varphi(p) = p - 1$, since every positive number less than p is relatively prime to p . In fact, if p is a prime and m is a positive integer, then $\varphi(p^m) = p^m - p^{m-1}$ since the only numbers between 1 and p^m that are not relatively prime to p^m are $p, 2p, 3p, 4p, \dots, (p^{m-1})p$, and there are clearly exactly p^{m-1} of those. As every number n can be factored into the form $p_1^{m_1} p_2^{m_2} \cdots p_k^{m_k}$ where p_1, p_2, \dots, p_k are distinct primes, it is now clear that the following lemma will eventually provide the complete answer.

Lemma 5.23 If m and n are relatively prime positive integers, then $\varphi(mn) = \varphi(m)\varphi(n)$.

Proof. Let $\zeta = \cos 2\pi/m + i \sin 2\pi/m$ and $\eta = \cos 2\pi/n + i \sin 2\pi/n$. It follows from Proposition 5.19 that $(k, m) = 1$ if and only if ζ^k is a primitive m -th root of unity. Hence we can interpret $\varphi(m)$ as the number of primitive m -th roots of unity. The proposition will be proved by demonstrating that all of the primitive mn -th roots of unity are obtained, without repetition, when an arbitrary primitive m -th root of unity is multiplied by an arbitrary primitive n -th root of unity.

We first dispose of the possible redundancies in this process. Thus, suppose that ζ^a and ζ^x are two primitive m -th roots of unity with $1 \leq a, x < m$, that η^b and η^y are two primitive n -th roots of unity with $1 \leq b, y < n$, and that $\zeta^a \eta^b = \zeta^x \eta^y$. Then $\zeta^{a-x} = \eta^{y-b}$ and consequently

$$\zeta^{(a-x)n} = (\zeta^{a-x})^n = (\eta^{y-b})^n = (\eta^n)^{y-b} = (1)^{y-b} = 1.$$

Since ζ is a primitive m -th root of unity it follows that m is a divisor of $(a-x)^n$. However, m and n are relatively prime, and so we may conclude that m is a divisor of

$(a - x)$ alone. Since both a and x are between 0 and m , it follows that $a = x$. A similar argument allows us to conclude that $b = y$. Thus it has been demonstrated that the set of all products of primitive m -th roots by primitive n -th roots contains exactly $\varphi(m)\varphi(n)$ distinct elements.

It follows from Proposition 5.20 that each of the products $\zeta^a \eta^b$, with ζ, η, a , and b as above, has order mn and is therefore a primitive mn -th root of unity. It therefore only remains to show that every primitive mn -th root is accounted for by this process.

Let α be any primitive mn -th root of unity, and let A and B be two integers such that

$$Am + Bn = 1. \quad (5.24)$$

Then clearly $\alpha = \alpha^{Am} \alpha^{Bn}$. Now $(\alpha^{Am})^n = (\alpha^{mn})^A = (1)^A = 1$, and so α^{Am} is an n -th root of unity. We know from Equation 5.24 that $(A, n) = 1$. It therefore follows from Proposition 5.19 that α^{Am} is in fact primitive, since

$$o(\alpha^{Am}) = \frac{mn}{(Am, mn)} = \frac{mn}{m(A, n)} = n.$$

A similar argument establishes that α^{Bn} is a primitive m -th root of unity, and so products of the primitive m -th roots of unity with the primitive n -th roots do indeed cover all the primitive mn -th roots, each exactly once. Thus, $\varphi(m)\varphi(n) = \varphi(mn)$. ■

We are now ready to derive an explicit formula for the number of positive integers that are both less than n and relatively prime to it. By Proposition 5.19, this is also equal to the number of primitive n -th roots of unity.

Theorem 5.25 If n is any number with prime factorization $p_1^{m_1} p_2^{m_2} \cdots p_k^{m_k}$, then

$$\varphi(n) = \prod_{i=1}^k (p_i^{m_i} - p_i^{m_i-1}).$$

Proof. By Lemma 5.23, it suffices to prove this theorem for the case where n is the power of a single prime, i.e., where there exist a prime number p and an integer m such that $n = p^m$. However, as was noted just prior to Lemma 5.23, $\varphi(p^m) = p^m - p^{m-1}$, and so the theorem now follows immediately. ■

For example, since $100 = 2^2 5^2$, it follows that $\varphi(100) = (2^2 - 2)(5^2 - 5) = 40$.

Private Key:	$n = pq, \quad \varphi(n) = (p-1)(q-1),$ $e \in \mathbb{Z}_{\varphi(n)}^*, \quad d \equiv e^{-1} \pmod{\varphi(n)};$
Public Key:	e and n ;
Encryption:	$c \equiv m^e \pmod{n};$
Decryption:	$c^d \equiv (m^e)^d \equiv m \pmod{\varphi(n)}.$

Table 5.2 RSA encryption

Number theory in general and Euler's φ -function in particular were considered to be pure, as opposed to applied, mathematics. In the 1970s, however, mathematicians found a very useful application of elementary number theory to encryption and decryption.

The purpose of encryption is to secure communications and we propose to do this by means of a method that relies on the mathematics of the previous section. Suppose Xenon has a multitude of customers and he wishes to establish secure communications with each of his customers. He consults Fermat who sells him his two latest primes p and q which Xenon hides as a tattoo on his body. Xenon also computes and displays on his website, as part of the Public Key, both the product $n = pq$ and a positive integer e such that $1 < e < \varphi(n)$ and $\gcd(e, \varphi(n)) = 1$.

And so when the client wishes to communicate something to Xenon, she translates the message into a number m with $1 < m < n$, which she promptly raises to the e -th power and sends to Xenon. If the message m^e reaches Xenon unmolested, all he has to do to decrypt the message is to raise m^e to the d -th power, where $d \equiv e^{-1} \pmod{\varphi(n)}$ (see Table 5.2).

If, however, the Red Baron intercepts this message, he still does not know the values of $\varphi(n)$ and since he, the Red Baron, does not know $\varphi(n)$, he cannot find d notwithstanding the fact that he does know e .

Exercises 5.4

1. Compute $\varphi(24)$, $\varphi(144)$, and $\varphi(1,000)$.
2. Prove that $\varphi(n)$ is even for $n > 2$.
3. For what values of n is $\varphi(n)$ a prime number? Justify your answer.
4. For what values of n is $\varphi(n)$ the power of a single prime number? Justify your answer.

5. True or false: there are an infinite number of integers n such that $\varphi(n) < 100$. Justify your answer.
6. Prove that if n is any positive integer, then

$$\sum_{d|n} \varphi(d) = n.$$

7. Let m and $n > 1$ be positive integers such that $\varphi(mn) = \varphi(m)$. Prove that $n = 2$ and m is odd.
8. Prove that if $g = (m, n)$, then $\varphi(mn) = g\varphi(m)\varphi(n)/\varphi(g)$.
9. Prove that if $d \mid n$, then $\varphi(d) \mid \varphi(n)$.

Chapter Summary

Having proved the Binomial Theorem, we used it to derive Fermat's Theorem for exponents in arithmetic modulo p , which effectively states that the nonzero elements of \mathbb{Z}_p are all $(p-1)$ -th roots of unity. This allowed us to extend the notion of order to modular arithmetic and we derived some new theorems regarding the orders of roots of unity which apply to both the complex and the modular ones.

Chapter Review Exercises

Mark the following true or false.

1. $\binom{n}{2} = n(n-1)/2$.
2. $\binom{8}{3} = 56$.
3. The number of pairs that can be formed by selecting two elements from the set $\{a, b, c, d\}$ is 8.
4. $12^7 \equiv 1 \pmod{17}$.
5. $(3+8)^{11} \equiv 3^{11} + 8^{11} \pmod{11}$.
6. $13^{90} \equiv 1 \pmod{31}$.
7. If $\alpha(\zeta) = 144$, then $\alpha(\zeta^{120}) = 102$.
8. $(x+y+z)^5 = x^5 + y^5 + z^5 + 5x^4y + 5xy^4 + 5x^4z + 5xz^4 + 5y^4z + 5yz^4 + 10x^2y^3 + 10x^3y^2 + 10x^2z^3 + 10x^3z^2 + 10y^2z^3 + 10y^3z^2$.
9. $\varphi(n) < n$ for all integers $n > 2$.
10. $\varphi(15) = 8$.

New Terms

binomial coefficient, 76	modular roots of unity, 87
Euler φ -function, 93	Pascal's Identity, 77
Fibonacci numbers, 84	Pascal's Triangle, 77
method of generating functions, 81	primitive root, 88
modular order, 87	

Supplementary Exercises

1. Write a computer script which computes the order of any element modulo p .
2. Write a computer script that evaluates $\binom{m}{n}$ for any two positive integers $m \geq n$.
3. Find $\lim_{n \rightarrow \infty} \varphi(n)$.
4. Let F_n be the Fibonacci number of Exercise 5.2.2.2. Prove that $(F_m, F_n) = F_{(m,n)}$ and investigate the question of which Fibonacci numbers are prime.
5. For which positive integers n does \mathbb{Z}_n have an element of order $\varphi(n)$?
6. For each positive integer n and $a \in \mathbb{Z}_n$ investigate the length of the (eventually) repeating segment of a, a^2, a^3, a^4, \dots .
7. Find some more Carmichael numbers.

Chapter 6



POLYNOMIALS OVER A FIELD

THE FOCUS now shifts to the topic of polynomials. Since polynomials have numerical coefficients, and since we have by now encountered a great variety of disparate number systems, the polynomials are studied in the more general setting of abstract fields. We are mainly concerned here with the factorization of polynomials in one variable, but some attention is also given to the symmetric polynomials in several variables and their utility in solving the general quartic equation.

6.1 Fields and Their Polynomials

We have by now encountered a host of algebraic structures within which the four arithmetical operations hold sway. These are the real numbers, the rational numbers, the complex numbers, and arithmetic modulo p where p is a prime. Another collection of such structures will be studied in detail in Chapter 7, and mathematicians have constructed many others that will not be mentioned here. It stands to reason that algebraic structures with such strong similarities will share yet other properties, and these are this chapter's concern. The notion of a field is used to extract the properties that are common to all these similar structures.

A *field* is a set F with two binary operations, usually denoted by $+$ and \cdot , for which the following hold. For any elements a , b , and c of F ,

$$\begin{array}{ll} a + b \in F \quad \text{and} \quad a \cdot b \in F & \text{(closure),} \\ (a + b) + c = a + (b + c) \quad \text{and} \quad (a \cdot b) \cdot c = a \cdot (b \cdot c) & \text{(associativity),} \\ a + b = b + a \quad \text{and} \quad a \cdot b = b \cdot a & \text{(commutativity),} \\ a \cdot (b + c) = a \cdot b + a \cdot c & \text{(distributivity),} \end{array}$$

there exist distinct elements $0, 1 \in F$ such that

$$a + 0 = a \quad \text{and} \quad a \cdot 1 = 1 \quad (\text{identities}),$$

there exists an element $-a \in F$ such that

$$a + (-a) = 0 \quad (\text{additive inverse}),$$

and if $a \neq 0$, then there exists an element $a^{-1} \in F$ such that

$$a \cdot a^{-1} = 1 \quad (\text{multiplicative inverse}).$$

It will prove useful to label the most familiar fields with some symbols. Accordingly, \mathbb{Q} , \mathbb{R} , and \mathbb{C} will denote the rational, real, and complex number systems, respectively. The fact that all these number systems are indeed fields will not be belabored here. It should be noted, however, that not all algebraic structures are necessarily fields. If n is a composite number, then \mathbb{Z}_n is not a field since, as was noted in Chapter 4, no divisor of n except 1 has a multiplicative inverse in \mathbb{Z}_n . The set of polynomials with real coefficients is another example of an algebraic structure that is not a field. It is easy to convince oneself that the normal addition and multiplication of such polynomials have all the properties required of a field, except for the last one. The multiplicative inverses of polynomials are not polynomials. For example, there is no polynomial whose product with $x + 1$ is 1 (Exercise 6.1.4).

Since the above-listed properties are shared by all fields, it is clear that any proposition whose justification relies only on these common properties is necessarily valid in all fields. In particular, it will hold for \mathbb{Q} , \mathbb{R} , \mathbb{C} , for all \mathbb{Z}_p with p prime, and for the new fields to be introduced in the next chapter. The following is an example of such a proposition.

Proposition 6.1 If a and b are two elements of the field F , then $a \cdot b = 0$ if and only if either a or b is zero.

Proof. Set $x = a \cdot 0$. By the distributivity of addition and multiplication,

$$x = a \cdot 0 = a \cdot (0 + 0) = a \cdot 0 + a \cdot 0 = x + x.$$

Consequently,

$$a \cdot 0 = x = x + (x + (-x)) = (x + x) + (-x) = x + (-x) = 0.$$

Conversely, suppose $a \cdot b = 0$ but $a \neq 0$. Then a has a multiplicative inverse a^{-1} and so

$$0 = a^{-1} \cdot 0 = a^{-1} \cdot (a \cdot b) = (a^{-1} \cdot a) \cdot b = 1 \cdot b = b.$$

Hence, either a or b is zero. ■

It is important to note that Proposition 6.1 does not hold in all algebraic structures. Thus, in \mathbb{Z}_6 , $2 \cdot 3 \equiv 0$ even though neither 2 nor 3 is zero. Another, more substantial example of a proposition that does hold for all fields is the Binomial Theorem. It was already stated in Section 5.1 that the proof of this theorem holds regardless of whether the numbers in question are complex or modular. In fact, the proof of Theorem 5.3 carries over verbatim to arbitrary fields once the meaning of the terms is clarified. For any positive integer m and any element a of some field F , let a^m denote the product of m a 's, and let ma denote the sum of m a 's. Accordingly, $a^3 = a \cdot a \cdot a$ and $3a = a + a + a$. If m is a negative integer we define $a^m = (a^{-1})^{-m}$ and $ma = -(-m)a$. Finally, we set $a^0 = 1$ and $0a = 0$. It is easily seen that such identities as $a^m \cdot a^n = a^{m+n}$ and $ma + na = (m+n)a$ hold in this generalized context just as they do for the complex and modular fields (Exercises 6.1.24 and 6.1.25).

Theorem 6.2 (The Binomial Theorem) Let F be a field, $a, b \in F$, and let n be a nonnegative integer. Then

$$(a + b)^n = \binom{n}{0}a^n + \binom{n}{1}a^{n-1}b + \cdots + \binom{n}{k}a^{n-k}b^k + \cdots + \binom{n}{n}b^n.$$

We now go on to show that many of the properties of polynomials with real coefficients are also valid when the coefficients are allowed to be the members of an arbitrary field. This will be accomplished by proving that the validity of these properties follows from the defining properties of fields listed above alone. For the sake of simplifying the notation we adopt here the usual convention that the product $a \cdot b$ is abbreviated to ab .

Given a field F , the *variables* x, y, z, \dots are symbols, or place holders, which can be replaced by elements of F . A *polynomial in x over F* is an expression that is obtained by applying the operations of addition and multiplication to the variable x and/or some of the elements of F . Thus, both $17x^3 - (31/2)x$ and $5x^2 + 6x + (-6)x^3 + 1 + (2/9)x$ are polynomials in x over \mathbb{Q} . Actually, they can also be interpreted as polynomials over \mathbb{R} , over \mathbb{C} , and over \mathbb{Z}_7 or any \mathbb{Z}_p (for $p \neq 2, 3$) for that matter. On the other hand,

$$x^3 + ix^2 + 1 - 3i$$

is a polynomial over the field of complex numbers \mathbb{C} , but not over either the real numbers or the rational numbers. Moreover, since in \mathbb{Z}_5

$$2^2 \equiv 3^2 \equiv 4 \equiv -1 \pmod{5}$$

it follows that in \mathbb{Z}_5 the quantity $i = \sqrt{-1}$ can be interpreted as either 2 or 3, and so $x^3 + ix^2 + 1 - 3i$ can be interpreted as a polynomial over \mathbb{Z}_5 . On the other hand, since \mathbb{Z}_3 contains no number a such that $a^2 \equiv 2 \equiv -1 \pmod{3}$, the polynomial $x^3 + ix^2 + 1 - 3i$ cannot be regarded as a polynomial over \mathbb{Z}_3 .

The set of polynomials in x over F is denoted by $F[x]$ and its members will generally be denoted by $P(x), Q(x), \dots$. If $P(x) \in F[x]$, then we also say that F is the *ground field* of $P(x)$. The two polynomials

$$P(x) = a_0x^n + a_1x^{n-1} + a_2x^{n-2} + \dots + a_{n-1}x + a_n$$

and

$$Q(x) = b_0x^m + b_1x^{m-1} + b_2x^{m-2} + \dots + b_{m-1}x + b_m$$

are said to be equal if and only if $m = n$ and $a_i = b_i$ for $i = 0, 1, 2, \dots, m = n$. In particular, the polynomials x^5 and x are considered to be distinct in $\mathbb{Z}_5[x]$ even though $n^5 \equiv n$ for all $n \in \mathbb{Z}_5$. The reason for this fine distinction will become clear in the next chapter.

Polynomials over an arbitrary field F can be added, subtracted, and multiplied, and these operations possess the usual properties of commutativity, associativity, and distributivity. Thus, the following two polynomials over \mathbb{Z}_5 are added and multiplied as

$$(2x^3 + 3x + 1) + (4x^3 + 2) = 6x^3 + 3x + 3 = x^3 + 3x + 3$$

and

$$\begin{aligned} (2x^3 + 3x + 1)(4x^3 + 2) &= 8x^6 + 4x^3 + 12x^4 + 6x + 4x^3 + 2 \\ &= 8x^6 + 12x^4 + 8x^3 + 6x + 2 = 3x^6 + 2x^4 + 3x^3 + x + 2. \end{aligned}$$

The details of these examples indicate that it is always necessary to keep the ground field in mind, since the final result clearly depends on which field the coefficients belong to. Many significant properties of a polynomial also depend on the ground field. The

polynomial $x^2 + 1$ factors over the complex numbers, since $x^2 + 1 = (x + i)(x - i)$, but it is well known that this polynomial cannot be factored over either the rationals or the real numbers. The same polynomial factors over \mathbb{Z}_5 as $x^2 + 1 = (x + 2)(x + 3)$, but does not factor over \mathbb{Z}_7 (see Proposition 6.6 below).

The polynomial 0 is called the *zero polynomial*. If $P(x)$ is any nonzero polynomial, then it can clearly be written in the standard form

$$a_0x^n + a_1x^{n-1} + a_2x^{n-2} + \cdots + a_{n-1}x + a_n, \quad a_0 \neq 0.$$

If $a_0 = 1$, the polynomial is said to be a *monic polynomial*. The exponent n is the *degree* of $P(x)$. No degree is assigned to the zero polynomial. A polynomial of degree 0 is said to be a *constant polynomial*. The zero polynomial is also considered to be a constant polynomial. The proof of the following proposition is straightforward and is relegated to Exercises 6.1.22 and 6.1.23.

Proposition 6.3 Let $P(x)$ and $Q(x)$ be polynomials of degrees m and n . Then if $P(x) + Q(x)$ is nonzero, then degree of $P(x) + Q(x)$ is at most $\max\{m, n\}$ and the degree of $P(x)Q(x)$ equals $m + n$.

We shall now address the issue of *division of polynomials* in $F[x]$. Much like the integers, polynomials are also subject to a process of long division. Because of the fundamental significance of this process, we shall prove its validity for polynomials over arbitrary ground fields. For the sake of completeness we next offer a proposition that justifies the process of long division in the general context of fields. The examples that follow the proof should clarify it and may actually obviate the need for such a proof. Given any two elements a and $b \neq 0$ of a field F , we use the symbol a/b to denote $a \cdot b^{-1}$ in the usual way.

Proposition 6.4 If $P(x)$ and $D(x) \neq 0$ are two polynomials over F , then there exist polynomials $Q(x), R(x) \in F[x]$ such that

$$P(x) = D(x)Q(x) + R(x) \tag{6.5}$$

and if $R(x)$ is not the zero polynomial, then the degree of $R(x)$ is less than the degree of $D(x)$.

Proof. Suppose

$$P(x) = a_0x^m + a_1x^{m-1} + \cdots + a_m$$

with $a_0 \neq 0$ and

$$D(x) = b_0x^d + b_1x^{d-1} + \cdots + b_d$$

with $b_0 \neq 0$. If $m < d$ then we can clearly choose $Q(x) = 0$ and $R(x) = P(x)$. Hence we may assume that $m \geq d$. We now proceed by induction on m and assume that the theorem holds for all pairs of polynomials $P(x)$, $D(x)$ with degrees less than some fixed integer $m \geq d$. Let $P(x)$ and $D(x)$ be as given and define the new polynomial

$$\begin{aligned} P_1(x) &= \frac{a_0}{b_0}x^{m-d}D(x) \\ &= \frac{a_0}{b_0}x^{m-d}(b_0x^d + b_1x^{d-1} + \cdots + b_d) \\ &= a_0x^m + \frac{a_0b_1}{b_0}x^{m-1} + \cdots + \frac{a_0b_d}{b_0}x^{m-d}. \end{aligned}$$

Then, because $P(x)$ and $P_1(x)$ have the same degree and the same first coefficient, it follows that $P(x) - P_1(x)$ is either 0 or else it is a polynomial of degree less than their common degree m . In the first case $P(x) = P_1(x)$ and we can use $Q(x) = (a_0/b_0)x^{m-d}$ and $R(x) = 0$ to obtain Equation 6.5. In the second case we use the induction hypothesis on the degrees. Accordingly, there exist polynomials $Q_1(x)$ and $R(x)$, with $R(x)$ either 0 or else of degree less than d , such that $P(x) - P_1(x) = Q_1(x)D(x) + R(x)$. But then

$$\begin{aligned} P(x) &= P_1(x) + Q_1(x)D(x) + R(x) = \frac{a_0}{b_0}x^{m-d}D(x) + Q_1(x)D(x) + R(x) \\ &= \left[\frac{a_0}{b_0}x^{m-d} + Q_1(x) \right] D(x) + R(x), \end{aligned}$$

and so, with $Q(x) = (a_0/b_0)x^{m-d} + Q_1(x)$, the proof is concluded. ■

The proof of the above proposition, like most inductive proofs, is in fact constructive and contains a method for finding the quotient $Q(x)$ and the remainder $R(x)$ of any long division of polynomials. We demonstrate this by dividing the polynomial $x^5 + x^4 + x^2 + 1$ by the polynomial $x^3 + x + 1$.

$$\begin{array}{r}
 x^2 + x + 1 \\
 x^3 + x + 1 \overline{) x^5 + x^4 + x^2 + 1} \\
 \underline{x^5 + x^3 + x^2} \\
 x^4 - x^3 \\
 \underline{x^4 + x^2 + x} \\
 -x^3 - x^2 - x + 1 \\
 \underline{-x^3 - x - 1} \\
 -x^2 + 2
 \end{array}$$

In this case, where the coefficients are real, $Q(x) = x^2 + x - 1$ and $R(x) = -x^2 + 2$. On the other hand, if the coefficients are taken as elements of \mathbb{Z}_7 , we get

$$\begin{array}{r}
 x^2 + x + 1 \\
 x^3 + x + 1 \overline{) x^5 + x^4 + x^2 + 1} \\
 \underline{x^5 + x^3 + x^2} \\
 x^4 + x^3 \\
 \underline{x^4 + x^2 + x} \\
 x^3 + x^2 + x + 1 \\
 \underline{x^3 + x + 1} \\
 x^2
 \end{array}$$

so that here $Q(x) = x^2 + x + 1$ and $R(x) = x^2$. Finally, if the division is undertaken over \mathbb{Z}_3 , we get

$$\begin{array}{r}
 \overline{x^2 + x + 1} \\
 x^3 + x + 1 \overline{) x^5 + x^4 + x^2 + 1} \\
 \underline{x^5 + x^3 + x^2} \\
 x^4 + 2x^3 + x \\
 \underline{x^4 + x^2 + x} \\
 2x^3 + 2x^2 + 2x + 1 \\
 \underline{2x^3 + 2x + 2} \\
 2x^2 + 2
 \end{array}$$

so that in this case $Q(x) = x^2 + x + 2$ and $R(x) = 2(x^2 + 1)$.

Exercises 6.1

1. Find the quotient and remainder when $x^7 + x^4 + x + 1$ is divided by $x^3 + x^2 + 1$ over \mathbb{Z}_2 .
2. Repeat Exercise 6.1.1 over \mathbb{Z}_3 .
3. Repeat Exercise 6.1.1 over \mathbb{Z}_5 .
4. Prove that there is no polynomial $P(x)$ with real coefficients such that $P(x)(x+1) = 1$.
5. Prove that if n is a composite number, then \mathbb{Z}_n has nonzero elements a and b such that $ab = 0$.

Prove that the identities in Exercises 6.1.6 to 6.1.9 hold in every field.

6. $(a+b)(a-b) = a^2 - b^2$
7. $(a+b)(a^2 - ab + b^2) = a^3 + b^3$
8. $(a-b)(a^2 + ab + b^2) = a^3 - b^3$
9. $1 + a + a^2 + \cdots + a^n = \frac{a^{n+1} - 1}{a - 1}$ if $a \neq 1$
10. Expand $(x+1)^6$ over \mathbb{Z}_2 .
11. Expand $(x+1)^6$ over \mathbb{Z}_3 .
12. Expand $(2x+3)^6$ over \mathbb{Z}_5 .
13. Expand $(2x+3)^6$ over \mathbb{Z}_7 .

Suppose a, b, c, d, e , and f are nonzero elements of field F such that

$$\frac{a}{b} = \frac{c}{d} = \frac{e}{f}.$$

Prove the identities in Exercises 6.1.14 to 6.1.17 whenever the denominators in question are nonzero.

14. $(a + b)/(a - b) = (c + d)/(c - d)$
15. $a/b = (a + c + e)/(b + d + f)$
16. $a/b = (a + 2c - 3e)/(b + 2d - 3f)$
17. $a/b = \sqrt[3]{(a^3 - 15c^3 + 9e^3)/(b^3 - 15d^3 + 9f^3)}$

Let F be any field and let a and x be elements of F . Prove the statements in Exercises 6.1.18 to 6.1.21.

18. If $a + x = a$, then $x = 0$.
20. If $a + x = 0$, then $x = -a$.
19. If $ax = a$ and $a \neq 0$, then $x = 1$.
21. If $ax = 1$, then $x = a^{-1}$.
22. Prove the first part of Proposition 6.3 and explain why equality need not hold.
23. Prove the second part of Proposition 6.3.
24. Prove that if m and n are positive integers and $a \in F$, then $a^m \cdot a^n = a^{m+n}$ and $(a^m)^n = a^{mn}$.
25. Prove that if m and n are integers and $a \in F$, then $ma + na = (m + n)a$ and $m(na) = (mn)a$.
26. Prove Exercise 6.1.24 when m and n are arbitrary integers.

6.2 The Factorization of Polynomials

If $P(x)$ is a polynomial in $F[x]$, and if a is an element of the field F such that $P(a) = 0$, then we say that a is a *zero* of $P(x)$. Thus,

- -3 is a zero of $2x + 6$ if $F = \mathbb{Q}$,
- 5 is a zero of $x^3 + 3x$ if $F = \mathbb{Z}_7$,
- $\sqrt{7}$ is a zero of $x^2 - 7$ if $F = \mathbb{R}$, and
- $1 + i$ is a zero of $x^4 + 4$ if $F = \mathbb{C}$.

Every precalculus algebra text contains a proposition that relates the zeros of a polynomial to its factorization. The same relationship holds for polynomials over arbitrary fields.

Proposition 6.6 Let $P(x)$ be a polynomial over the field F . Then $a \in F$ is a zero of $P(x)$ if and only if there exists a polynomial $Q(x)$ such that $P(x) = Q(x)(x - a)$.

Proof. If such a polynomial $Q(x)$ does exist, then clearly $P(a) = Q(a)(a - a) = 0$, and so a is indeed a zero of $P(x)$. Conversely, suppose a is a zero of $P(x)$. Set $D(x) = x - a$, and let $Q(x)$ and $R(x)$ be the polynomials whose existence is guaranteed by Proposition 6.4. Then $P(x) = Q(x)(x - a) + R(x)$. If $R(x)$ is the zero polynomial, we are done. Otherwise the degree of $R(x)$ is less than that of $D(x) = x - a$ which is 1. Hence $R(x)$ is a polynomial of degree zero, so that for some $r \in F$, $P(x) = Q(x)(x - a) + r$. If we now substitute $x = a$, we obtain $0 = P(a) = Q(a)(a - a) + r = 0 + r = r$, so that $P(x) = Q(x)(x - a)$. ■

This proposition has a corollary whose straightforward inductive proof is relegated to Exercise 6.2.22.

Corollary 6.7 If $P(x)$ is a polynomial over the field F and if $a_1, a_2, \dots, a_k \in F$ are distinct zeroes of $P(x)$, then there exists a polynomial $Q(x)$ such that

$$P(x) = Q(x)(x - a_1)(x - a_2) \cdots (x - a_k).$$

If $P(x)$, $Q(x)$, and $R(x)$ are polynomials over F such that $P(x) = Q(x)R(x)$, then we say that $Q(x)R(x)$ is a *factorization* of $P(x)$ over F , and that $Q(x)$ and $R(x)$ are *factors* or *divisors* of $P(x)$. If $P(x)$ is a nonconstant polynomial such that in every factorization $P(x) = Q(x)R(x)$ either $Q(x)$ or $R(x)$ has degree 0 (i.e., is a nonzero constant), then $P(x)$ is said to be *irreducible* over F . A polynomial that is not irreducible is called *reducible* or *factorable*.

The above proposition greatly facilitates the task of factoring polynomials. The polynomial $P_1(x) = x^2 + x + 3$ has the two zeroes 1 and 3 over \mathbb{Z}_5 since $P_1(1) = 5 \equiv 0 \pmod{5}$ and $P_1(3) = 15 \equiv 0 \pmod{5}$. Hence it follows from Proposition 6.6 that $(x - 1) = (x + 4)$ and $(x - 3) = (x + 2)$ are factors of $x^2 + x + 3$ (over \mathbb{Z}_5). Since $x^2 + x + 3$ has degree 2, it can have at most two factors. Consequently, by Corollary 6.7, $x^2 + x + 3 = (x + 2)(x + 4)$ over \mathbb{Z}_5 .

Similarly, the polynomial $P_2(x) = 2x^2 + 2x + 3$ has the two zeroes 1 and 5 (mod 7). Thus $(x - 1)(x - 5) = (x + 6)(x + 2)$ is a divisor of $P_2(x)$. Since the leading coefficient of $P_2(x)$ is 2 it follows that $2x^2 + 2x + 3 = 2(x + 2)(x + 6)$ over \mathbb{Z}_7 .

The polynomial $P_3(x) = x^2 + x + 1$ has no zeroes in \mathbb{Z}_2 since $P_3(0) = 1 \not\equiv 0 \pmod{2}$ and $P_3(1) = 3 \not\equiv 0 \pmod{2}$. Since every nonconstant factor of $P_3(x)$ would have form $x - a$, $a \in \{0, 1\}$, and since neither 0 nor 1 are zeros of $P_3(x)$, it now follows from

Proposition 6.6 that this polynomial is irreducible over \mathbb{Z}_2 . On the other hand, since $P_3(1) \equiv 0 \pmod{3}$, it follows that $P_3(x)$ does factor over \mathbb{Z}_3 in a nontrivial way.

It is easy to read too much into Corollary 6.7. The polynomial $x^4 + x^2 + 1$ has no zeroes in \mathbb{Z}_2 for the same reasons that $P_3(x)$ does not. Nevertheless,

$$\begin{aligned}(x^2 + x + 1)^2 &= (x^2 + x + 1)(x^2 + x + 1) \\ &= x^4 + x^3 + x^2 + x^3 + x^2 + x + x^2 + x + 1 = x^4 + x^2 + 1.\end{aligned}$$

Thus, Corollary 6.7 only supplies information about first-degree factors. It may fail to detect the existence of factors of a higher degree.

The question of which polynomials of degree greater than 3 are irreducible is quite difficult and cannot be discussed here in its full generality. In the case where the ground field is \mathbb{Z}_p , however, the finiteness of p can be used to make some headway. Thus, the irreducible polynomials of degree 1, 2, or 3 over \mathbb{Z}_2 are (Exercise 6.2.1) x , $x + 1$, $x^2 + x + 1$, $x^3 + x + 1$, and $x^3 + x^2 + 1$. Hence the list of reducible fourth-degree polynomials over \mathbb{Z}_2 is

$$\begin{array}{ll}x^4, & (x + 1)^2(x^2 + x + 1) = x^4 + x^3 + x + 1, \\x^3(x + 1) = x^4 + x^3, & (x^2 + x + 1)^2 = x^4 + x^2 + 1, \\x^2(x + 1)^2 = x^4 + x^2, & x(x^3 + x + 1) = x^4 + x^2 + x, \\x(x + 1)^3 = x^4 + x^3 + x^2 + x, & x(x^3 + x^2 + 1) = x^4 + x^3 + x, \\(x + 1)^4 = x^4 + 1, & (x + 1)(x^3 + x + 1) = x^4 + x^3 + x^2 + 1, \\x^2(x^2 + x + 1) = x^4 + x^3 + x^2, & (x + 1)(x^3 + x^2 + 1) = x^4 + x^2 + x + 1, \\x(x + 1)(x^2 + x + 1) = x^4 + x, & \end{array}$$

The remaining three fourth-degree polynomials over \mathbb{Z}_2 , namely, $x^4 + x^3 + 1$, $x^4 + x + 1$, and $x^4 + x^3 + x^2 + x + 1$, must therefore be irreducible.

It is important to realize that the same polynomial may be factorable over one field and irreducible over another. Thus, we saw above that $x^2 + x + 1$ is irreducible over \mathbb{Z}_2 whereas it is easily verified that $x^2 + x + 1 = (x + 2)^2$ over \mathbb{Z}_3 . Similarly, the polynomial $x^2 + 1$ is irreducible over \mathbb{R} (it has degree 2 and no zeroes in \mathbb{R}) whereas it factors into $(x + i)(x - i)$ over \mathbb{C} .

We conclude this section by extending to arbitrary fields yet another fact that is well known for polynomials over the real numbers.

Proposition 6.8 If $P(x)$ is a polynomial of degree n over any field F , then the equation $P(x) = 0$ has at most n distinct solutions.

Proof. This is proved by induction on n . When $n = 0$, $P(x)$ must be a constant polynomial c for some $c \neq 0$. In this case the polynomial equation $P(x) = 0$ has the form $c = 0$ which has no (i.e., zero) solutions. Hence the induction process has been anchored at $n = 0$.

Let $P(x)$ be a polynomial of degree $n > 0$ and suppose that the theorem has been proved for all polynomials of degree less than n . If $P(x)$ has no zeroes, then we are done. Suppose, therefore, that $r \in F$ is a zero of $P(x)$. Proposition 6.6 implies the existence of a polynomial $Q(x)$ over F such that $P(x) = Q(x)(x - r)$. It is clear that $Q(x)$ has degree $n - 1$. If s is any zero of $P(x)$ that is distinct from r , then $0 = P(s) = Q(s)(s - r)$. Since s is distinct from r , it follows that $s - r \neq 0$ and hence, by Proposition 6.1, we may conclude that $Q(s) = 0$. Thus, all the zeroes of $P(x)$ that are distinct from r are zeroes of $Q(x)$. Since $Q(x)$ has degree $n - 1$, there are, by the induction hypothesis, at most $n - 1$ zeroes of $P(x)$ that are distinct from r . In other words, $P(x)$ has at most n distinct zeroes. ■

The actual number of distinct solutions of a polynomial equation will vary with the polynomial. It is easily verified by direct substitution that the equation

$$x^3 + x^2 + x + 1 \equiv 0 \pmod{5}$$

has the solutions $x = 2$, $x = 3$, and $x = 4$ in \mathbb{Z}_5 , whereas the equation

$$x^3 + 3x + 4 \equiv 0 \pmod{5}$$

has only two distinct solutions, namely, $x = 3$ and $x = 4$. However, since over \mathbb{Z}_5 we have $x^3 + 3x + 4 = (x - 3)^2(x - 4)$, we say that 3 is a double zero of the polynomial $x^3 + 3x + 4$, and consequently, counting multiplicities, this polynomial has three zeroes. In general, if r is any zero of the polynomial $P(x)$, it is said to have *multiplicity* m if m is the largest integer such that $(x - r)^m$ divides $P(x)$.

Thus, as was seen above, 3 and 4 are zeroes of multiplicities 2 and 1, respectively, of the polynomial $x^3 + 3x + 4$ over \mathbb{Z}_5 . A slight modification of the proof of Proposition 6.8, together with the Fundamental Theorem of Algebra that was stated without proof in Section 3.3, yields the following fact whose proof is relegated to Exercise 6.2.20.

Proposition 6.9 Counting multiplicities, every polynomial of degree n over the complex numbers has exactly n complex zeroes.

This proposition is in some sense also true even when $P(x)$ has its coefficients in other fields (Exercise 10.3.23).

Amongst the consequences of the Fundamental Theorem of Algebra are the following corollaries:

Lemma 6.10 If $P(x)$ is a polynomial with real coefficients, then r is a zero of $P(x)$ if and only if \bar{r} is one too.

Proof. Let

$$P(x) = a_0x^k + a_1x^{k-1} + a_2x^{k-2} + \cdots + a_{k-1}x + a_k.$$

Because the coefficients a_0, a_1, \dots, a_k are real, it follows that

$$\overline{P(x)} = \overline{\sum_{i=0}^k a_i x^{k-i}} = \sum_{i=0}^k \overline{a_i} \overline{x}^{k-i} = \sum_{i=0}^k a_i \bar{x}^{k-i} = P(\bar{x}).$$

Consequently r is a zero of $P(x)$ if and only if \bar{r} is such a zero too. ■

Corollary 6.11 If $P(x)$ is a polynomial with real coefficients, then it can be factored into irreducible real polynomials of degree at most 2.

Proof. Let $P(x) = a_0x^k + a_1x^{k-1} + a_2x^{k-2} + \cdots + a_{k-1}x + a_k$. By Proposition 6.9 there exist complex numbers r_1, r_2, \dots, r_k such that

$$P(x) = a_0 \prod_{i=1}^k (x - r_i).$$

By the above lemma, if $P(x)$ has m real zeros, then the number of imaginary zeros is $(k-m)/2$ and $P(x)$ factors into

$$a_0 \prod_{i=1}^m (x - r_i) \prod_{j=1}^{(k-m)/2} (x - s_j)(x - \bar{s}_j)$$

where each r_i is real and each s_j is a nonreal imaginary number. However,

$$(x - s_j)(x - \bar{s}_j) = x^2 - (s_j + \bar{s}_j)x + s_j \bar{s}_j$$

where both $s_j + \bar{s}_j$ and $s_j \bar{s}_j$ are real. Thus, $P(x)$ has been factored into linear and quadratic irreducible factors. ■

Exercises 6.2

1. List and completely factor all the polynomials of degree $d \leq 4$ over \mathbb{Z}_2 .
2. List and completely factor all the polynomials of the form $x^5 + ax^2 + bx + c$ over \mathbb{Z}_2 .
3. List and completely factor all the monic quadratic polynomials over \mathbb{Z}_3 .
4. List and completely factor all the cubic polynomials of the form $x^3 + x^2 + ax + b$ over \mathbb{Z}_3 .
5. Completely factor all the polynomials of the form $x^3 + ax + 1$ over \mathbb{Z}_5 .
6. Prove that the number of irreducible monic quadratic polynomials over \mathbb{Z}_p is $\binom{p}{2}$.
7. Find a formula for the number of irreducible monic cubic polynomials over \mathbb{Z}_p .
8. Prove that if F is a finite field, then there is a quadratic polynomial in $F[x]$ that is irreducible over F .
9. Suppose the polynomial $a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n$ is irreducible over a field F , with $a_0, a_n \neq 0$. Prove that the polynomial $a_nx^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$ is also irreducible over F .
10. Prove that if the polynomial $P(x) \in F[x]$ is divided by $x - a$, then the remainder is $P(a)$.
11. Let a be any element of the field F , and let $P(x)$ be a polynomial over F . Prove that $P(x)$ is irreducible over F if and only if $P(x + a)$ is irreducible over F .

Find the remainder when $P(x) = x^{25} - 2x^4 + 3x^3 - 4x^2 + 5x - 1 \in \mathbb{C}[x]$ is divided by the polynomials in Exercises 6.2.12 to 6.2.15.

12. $x - 1$
13. $x^2 - x$
14. $x^2 - 1$
15. $x^2 + 1$
16. For which values of a and b will $x^2 + 2$ be a factor of $x^{17} + ax + b$ over \mathbb{Z}_3 ?
17. Repeat Exercise 6.2.16 over \mathbb{Z}_{11} .
18. Let F be an arbitrary field. Do there exist $a, b \in F$ such that $x^2 + 2$ is a divisor of $x^{17} + ax + b$ over F ?
19. Repeat Exercise 6.2.16 over \mathbb{R} .
20. Repeat Exercise 6.2.16 over \mathbb{C} .
21. Prove Proposition 6.9 assuming the Fundamental Theorem of Algebra (Section 3.3).
22. Prove Corollary 6.7.
23. Prove that the polynomial $x^4 + x^3 + x^2 + x + 1$ is irreducible over \mathbb{Q} .

$$\begin{array}{c}
 M_{i-1}(x) = Q_{i-1}(x)N_{i-1}(x) = R_{i-1}(x) \\
 \swarrow \quad \searrow \\
 M_i(x) = Q_i(x)N_i(x) = R_i(x) \\
 \swarrow \quad \searrow \\
 M_{i+1}(x) = Q_{i+1}(x)N_{i+1}(x) = R_{i+1}(x)
 \end{array}$$

Figure 6.1 The Euclidean algorithm for polynomials

6.3 The Euclidean Algorithm for Polynomials

The process of long division was used once before in Chapter 4 in connection with the Euclidean algorithm for finding the greatest common divisor of two integers. It proves equally handy in finding the greatest common divisor of two polynomials. We define a *greatest common divisor* of the polynomials $M(x)$ and $N(x)$ as a common divisor of maximum degree. The reason for the use of the indefinite article in this definition is that if the polynomial $P(x)$ is a divisor of $Q(x)$ and c is any nonzero number in the ground field, then so is $cP(x)$ a divisor of $Q(x)$. For if $Q(x) = P(x)R(x)$, then also $Q(x) = [cP(x)][(1/c)R(x)]$. Thus, if $G(x)$ is any greatest common divisor of two polynomials, then so is $cG(x)$ whenever $c \neq 0$. Exercise 6.3.8 calls for a proof that any two greatest common divisors of $M(x)$ and $N(x)$ are indeed such multiples of each other, but until then some caution must be exercised.

To find a greatest common divisor of two given polynomials it suffices to mimic the Euclidean algorithm for integers. We begin with a lemma which implies that the long division process of Proposition 6.4 can be used to reduce the degrees of the polynomials in question.

Lemma 6.12 Suppose $P(x)$, $Q(x)$, $D(x)$, and $R(x)$ are polynomials over the field F such that $P(x) = D(x)Q(x) + R(x)$, then every greatest common divisor of $P(x)$ and $D(x)$ is also a greatest common divisor of $D(x)$ and $R(x)$ and vice versa.

Proof. Every common divisor of $D(x)$ and $R(x)$ is also a divisor of $D(x)P(x) + R(x) = P(x)$, and hence it is also a common divisor of $D(x)$ and $P(x)$. Conversely, every common divisor of $P(x)$ and $D(x)$ is also a divisor of $P(x) - D(x)Q(x) = R(x)$, and hence it is also a common divisor of $D(x)$ and $R(x)$. The complete statement of the lemma now follows immediately. ■

Let $M(x)$ and $N(x)$ be two polynomials. Set $M_1(x) = M(x)$ and $N_1(x) = N(x)$ and let $Q_1(x)$ and $R_1(x)$ be the appropriate quotient and remainder so that

$$M_1(x) = Q_1(x)N_1(x) + R_1(x).$$

For $i = 1, 2, 3, \dots$ we set

$$M_{i+1}(x) = N_i(x) \quad \text{and} \quad N_{i+1}(x) = R_i(x) \quad (6.13)$$

with $Q_{i+1}(x)$ and $R_{i+1}(x)$ being the respective quotient and remainder when $M_{i+1}(x)$ is divided by $N_{i+1}(x)$. Figure 6.1 should be helpful.

Note that if $R_i(x)$ is not the zero polynomial, then either $R_{i+1}(x)$ is the zero polynomial or else

$$\text{degree of } R_{i+1}(x) < \text{degree of } N_{i+1}(x) = \text{degree of } R_i(x).$$

Hence, this procedure is bound to eventually produce a remainder $R_k(x)$ which is the zero polynomial, at which point the algorithm stops. We claim that $R_{k-1}(x)$ is a greatest common divisor of $M(x)$ and $N(x)$. To see this first note that the bottom line of Figure 6.1 is equivalent to

$$R_{i-1}(x) = Q_{i+1}(x)R_i(x) + R_{i+1}(x) \quad i = 2, 3, \dots, k-1.$$

Hence, by the above lemma, for $i = 2, 3, \dots, k-1$ any greatest common divisor of $R_{i-1}(x)$ and $R_i(x)$ is also a greatest common divisor of $R_i(x)$ and $R_{i+1}(x)$ and vice versa.

A similar argument allows us to conclude that any greatest common divisor of $R_2(x)$ and $R_1(x)$ is also a greatest common divisor of $M(x) = M_1(x)$ and $N(x) = N_1(x)$ (Exercise 6.3.20). Hence, since $R_{k-1}(x)$ is a greatest common divisor of $R_{k-1}(x)$ and $R_k(x)$, it is also a greatest common divisor of $M(x)$ and $N(x)$.

If the ground field is \mathbb{Z}_2 and

$$M(x) = M_1(x) = x^8 + x^7 + x^6 + x^4 + x^3 + x + 1$$

and

$$N(x) = N_1(x) = x^5 + x^4 + x^3 + x^2 + x + 1,$$

then two long divisions yield

$$M_2(x) = N_1(x) = x^5 + x^4 + x^3 + x^2 + x + 1,$$

$$N_2(x) = R_1(x) = x^4 + x^3 + x^2,$$

$$M_3(x) = N_2(x) = x^4 + x^3 + x^2,$$

$$N_3(x) = R_2(x) = x^2 + x + 1,$$

and $R_3(x) = 0$. Thus $R_2(x) = x^2 + x + 1$ is the required greatest common divisor of $M(x)$ and $N(x)$.

It will subsequently prove useful to have a polynomial version of Proposition 4.1 available.

Proposition 6.14 If $G(x)$ is a greatest common divisor of the polynomials $M(x)$ and $N(x)$ over the field F , then there exist polynomials $A(x)$ and $B(x)$ over F such that $A(x)M(x) + B(x)N(x) = G(x)$.

Proof. The proof we give applies only to the greatest common divisor obtained by the Euclidean algorithm, which we shall call the *Euclidean greatest common divisor*. The proposition's validity for all greatest common divisors then follows from Exercises 6.3.7 and 6.3.8.

We mimic the inductive proof of Proposition 4.1 and use the notation employed above in the description of the Euclidean algorithm for polynomials. Let k be the number of divisions in the application of the Euclidean algorithm to $M(x)$ and $N(x)$. If $k = 1$, this means that $R_1(x)$ is the zero polynomial so that $N(x)$ is a divisor of $M(x)$ and the Euclidean greatest common divisor $G(x)$ of $M(x)$ and $N(x)$ is $N(x)$ itself. Thus, choosing $A(x) = 0$ and $B(x) = 1$ we get $0 \cdot M(x) + 1 \cdot N(x) = G(x)$. Assume that the theorem holds for all pairs of polynomials for which the Euclidean algorithm requires $k - 1$ divisions. If $M(x)$ and $N(x)$ are a pair that require k divisions to arrive at their Euclidean greatest common divisor $G(x)$, then the pair $M_2(x)$ and $N_2(x)$ given by Equation 6.13 requires only $k - 1$ divisions to arrive at their Euclidean greatest common divisor which is also $G(x)$. By the induction hypothesis, there exist polynomials $A'(x)$ and $B'(x)$ such that

$$A'(x)M_2(x) + B'(x)N_2(x) = G(x).$$

However, if $Q_1(x)$ and $R_1(x)$ are the quotient and remainder obtained when $M(x)$ is divided by $N(x)$, then $M_2(x) = N_1(x)$ and $N_2(x) = R_1(x) = M_1(x) - Q_1(x)N_1(x)$, and

so

$$A'(x)N_1(x) + B'(x)[M_1(x) - Q_1(x)N_1(x)] = G(x),$$

or

$$B'(x)M(x) + [A'(x) - B'(x)Q_1(x)]N(x) = G(x),$$

so that $A(x) = B'(x)$ and $B(x) = A'(x) - B'(x)Q_1(x)$ are the required polynomials. ■

Consider $x^4 + x^3 + x + 1$ and $x^5 + x^2 + x + 1$ as polynomials in $\mathbb{Z}_2[x]$. Then

$$x^5 + x^2 + x + 1 = (x + 1)(x^4 + x^3 + x + 1) + (x^3 + x),$$

$$x^4 + x^3 + x + 1 = (x + 1)(x^3 + x) + (x^2 + 1),$$

and $x^3 + x = x(x^2 + 1)$. Hence, $x^2 + 1$ is a greatest common divisor of the two given polynomials, and

$$\begin{aligned} x^2 + 1 &= (x^4 + x^3 + x + 1) + (x + 1)(x^3 + x) \\ &= (x^4 + x^3 + x + 1) + (x + 1)[(x^5 + x^2 + x + 1) + (x + 1)(x^4 + x^3 + x + 1)] \\ &= [1 + (x + 1)^2](x^4 + x^3 + x + 1) + (x + 1)(x^5 + x^2 + x + 1) \\ &= x^2(x^4 + x^3 + x + 1) + (x + 1)(x^5 + x^2 + x + 1). \end{aligned}$$

Irreducible polynomials are the analogs of prime numbers, and just like the integers, polynomials have a unique factorization theorem. The formal statement of this theorem and its proof are relegated to Exercises 6.3.15 and 6.3.16.

Two polynomials are said to be *relatively prime* if their only common divisors are constants (i.e., elements of the ground field F). It is clear that such relatively prime polynomials have 1 as their greatest common divisor. Consequently we have the following proposition which will turn out to be very useful, in a nonobvious way, in Section 7.1.

Proposition 6.15 If $M(x)$ and $N(x)$ are relatively prime polynomials over a field F , then there exist polynomials $A(x), B(x) \in F[x]$ such that $A(x)M(x) + B(x)N(x) = 1$.

Exercises 6.3

1. Find a greatest common divisor of the polynomials $x^5 + x + 1$ and $x^6 + x^5 + x^4 + x^3 + 1$ over \mathbb{Z}_2 .
2. Find a greatest common divisor of the polynomials $x^6 + x^4 + x + 1$ and $x^7 + x^4 + x^3 + 1$ over \mathbb{Z}_2 .
3. Find a greatest common divisor of the polynomials $x^5 + x^4 + 2x^3 + x + 2$ and $x^6 + 2x^4 + x^2 + 2$ over \mathbb{Z}_3 .
4. Find a greatest common divisor of the polynomials $x^6 + x^5 + x^4 + 2$ and $x^7 + x^6 + x^5 + 2x^3 + x^2 + 2x + 1$ over \mathbb{Z}_3 .
5. Repeat Exercise 6.3.1 over \mathbb{Z}_5 .
6. Repeat Exercise 6.3.3 over \mathbb{Z}_5 .
7. Let $M(x)$ and $N(x)$ be any two polynomials over an arbitrary field F , and suppose $D(x)$ is another such polynomial that divides both $M(x)$ and $N(x)$. Prove that $D(x)$ divides the Euclidean greatest common divisor of $M(x)$ and $N(x)$.
8. Let $G(x)$ be the Euclidean greatest common divisor of $M(x), N(x) \in F[x]$, and let $H(x)$ be any polynomial over F . Prove that $H(x)$ is a greatest common divisor of $M(x)$ and $N(x)$ if and only if there exists a nonzero constant $c \in F$ such that $H(x) = cG(x)$.
9. Do there exist polynomials $A(x)$ and $B(x)$ over \mathbb{R} such that

$$A(x)(x^2 + 3x + 2) + B(x)(x^2 - 1) = x + 2?$$

10. Do there exist polynomials $A(x)$ and $B(x)$ over \mathbb{R} such that

$$A(x)(x^2 + 3x + 2) + B(x)(x^2 - 1) = x^2 - x - 2?$$

11. Do there exist polynomials $A(x)$ and $B(x)$ over \mathbb{R} such that

$$A(x)(x^2 - 5x + 6) + B(x)(x^2 + x - 6) = x^2 + 1?$$

12. Do there exist polynomials $A(x)$ and $B(x)$ over \mathbb{R} such that

$$A(x)(x^2 - 4) + B(x)(x^2 + 2x - 8) = x^2 - 2x?$$

13. Let $M(x)$ and $N(x)$ be any two polynomials over an arbitrary field F . Characterize all the polynomials K that can be expressed in the form

$$K(x) = A(x)M(x) + B(x)N(x)$$

for some polynomials $A(x)$ and $B(x)$ over F .

14. Let $M(x)$ and $N(x)$ be any two polynomials over an arbitrary field F . Of all the polynomials that can be expressed in the form $A(x)M(x) + B(x)N(x)$ for some $A(x), B(x) \in F[x]$, let $H(x)$ be one that possesses minimum degree. Prove that $H(x)$ is a greatest common divisor of $M(x)$ and $N(x)$.
15. Let $P(x), M(x), N(x)$ be polynomials over F such that $P(x)$ is irreducible and $P(x)$ is a divisor of the product $M(x)N(x)$. Prove that $P(x)$ is a divisor of either $M(x)$ or $N(x)$.
16. Let $N(x)$ be any monic polynomial over the field F . Prove that there exist monic irreducible polynomials $P_1(x), P_2(x), \dots, P_b(x)$ and positive integers r_1, r_2, \dots, r_b such that

$$N(x) = P_1^{r_1}(x)P_2^{r_2}(x) \cdots P_b^{r_b}(x).$$

Moreover, if $R_1(x), R_2(x), \dots, R_k(x)$ is another set of irreducible polynomials, and s_1, s_2, \dots, s_k is another set of positive integers such that

$$N(x) = R_1^{s_1}(x)R_2^{s_2}(x) \cdots R_k^{s_k}(x),$$

then $b = k$, and the R_i 's be reindexed so that $P_i = R_i$ and $r_i = s_i$ for $i = 1, 2, \dots, b$. (Hint: see Theorem 4.9.)

17. Suppose the polynomials $x^3 + 3px^2 + 3qx + r$ and $x^2 + 2px + q$ have a nonconstant greatest common divisor. Show that

$$4(p^2 - q)(q^2 - pr) - (pq - r)^2 = 0.$$

18. Prove that the polynomial $P(x) \in \mathbb{C}[x]$ has a zero of multiplicity at least two if and only if $(P(x), P'(x))$ is not a constant, where $P'(x)$ is the derivative of $P(x)$.
19. Prove that if the polynomial $ax^3 + 3bx^2 + 3cx + d \in \mathbb{C}[x]$ has a zero of multiplicity 2, then this zero is

$$\frac{bc - ad}{2(ac - b^2)}.$$

20. Supply the missing details in the verification of the Euclidean algorithm for polynomials by proving that any greatest common divisor of $R_2(x)$ and $R_1(x)$ is also a greatest common divisor of $M(x)$ and $N(x)$.
21. Let ζ be a complex primitive n -th root of unity. Prove that

$$\prod_{k=1}^n (x - \zeta^k) = x^n - 1.$$

22. Let a be any primitive root modulo p for some prime p . Prove that

$$\prod_{k=1}^{p-1} (x - a^k) = x^{p-1} - 1 \quad \text{over } \mathbb{Z}_p.$$

23. For each positive integer n let $\zeta_{1,n}, \zeta_{2,n}, \dots, \zeta_{m,n}$ be all the complex primitive n -th roots of unity (m will depend on n). Prove that the polynomial

$$P_n(x) = \prod_{k=1}^m (x - \zeta_{k,n})$$

has real rational coefficients for each n .

24. Prove that if $P(x)$ and $Q(x)$ are polynomials over both the fields $F \subset F'$, then the greatest common divisor of $P(x)$ and $Q(x)$ over F is also their greatest common divisor over F' .
25. The polynomials $x + 1$ and 1 are greatest common divisors of $x + 1$ and $x^2 + 1$ over \mathbb{Z}_2 and \mathbb{Z}_3 , respectively. Explain why this does not contradict Exercise 6.3.24.
26. Let $M(x) = x^5 + x + 1$ and $N(x) = x^6 + x^5 + x^4 + x^3 + 1$ be the polynomials of Exercise 6.3.1, and let $G(x)$ be their greatest common divisor over \mathbb{Z}_2 . Find polynomials $A(x)$ and $B(x)$ such that $G(x) = A(x)M(x) + B(x)N(x)$ over \mathbb{Z}_2 .
27. Let $M(x) = x^6 + x^4 + x + 1$ and $N(x) = x^7 + x^4 + x^3 + 1$ be the polynomials of Exercise 6.3.2, and let $G(x)$ be their greatest common divisor over \mathbb{Z}_2 . Find polynomials $A(x)$ and $B(x)$ such that $G(x) = A(x)M(x) + B(x)N(x)$ over \mathbb{Z}_2 .

6.4 Elementary Symmetric Polynomials

It was already observed in Chapter 1 that there is a close relationship between the coefficients of a quadratic equation and its roots. Namely, if r and s are the roots of the quadratic equation $ax^2 + bx + c = 0$, then

$$r + s = -\frac{b}{a} \quad \text{and} \quad rs = \frac{c}{a}.$$

This will now be generalized to polynomial equations of arbitrary degrees and with coefficients in arbitrary fields. First, however, we simplify the statements of the subsequent theorems by restricting attention to monic polynomials. Thus, for the monic quadratic equation $x^2 + bx + c = 0$, we have, by Proposition 1.3, $r + s = -b$ and $rs = c$. In general, suppose we have a monic polynomial

$$P(x) = x^n + a_1x^{n-1} + a_2x^{n-2} + \cdots + a_{n-1}x + a_n, \quad (6.16)$$

with coefficients in an arbitrary field F . Suppose further that $P(x)$ has been factored into linear factors so that

$$P(x) = (x - r_1)(x - r_2) \cdots (x - r_i) \cdots (x - r_n). \quad (6.17)$$

Then if all the $(x - r_i)$ of Equation 6.17 are multiplied out, and if all like terms are added, the right-hand side of Equation 6.16 should be obtained. Before these summands are added, there are 2^n of them, each having the form

$$(-1)^k A_1 A_2 \cdots A_i \cdots A_n, \quad (6.18)$$

where each A_i is either x or r_i , and k is the number of the A_i 's that equal r_i . For example, there are n summands that contain $n - 1$ x 's, namely

$$(-1)^1 r_1 x x x \cdots x x = -r_1 x^{n-1},$$

$$(-1)^1 x r_2 x x \cdots x x = -r_2 x^{n-1},$$

$$(-1)^1 x x r_3 x \cdots x x = -r_3 x^{n-1},$$

$$\vdots$$

$$(-1)^1 x x x x \cdots x r_n = -r_n x^{n-1}.$$

The sum of these terms must agree with the x^{n-1} term in the right-hand side of Equation 6.16 and so we conclude that

$$-r_1 x^{n-1} - r_2 x^{n-1} - r_3 x^{n-1} - \cdots - r_n x^{n-1} = a_1 x^{n-1},$$

or $r_1 + r_2 + r_3 + \cdots + r_n = -a_1$. Similarly, each summand in Equation 6.18 that contains $n-2$ x 's has two of its A_i 's equal to the corresponding r_i 's, the rest of the A_i 's being x , and the value of k is 2. Thus, we conclude that

$$r_1 r_2 x^{n-2} + r_1 r_3 x^{n-2} + \cdots + r_1 r_n x^{n-2} + r_2 r_3 x^{n-2} + \cdots + r_{n-1} r_n x^{n-2} = a_2 x^{n-2},$$

or $r_1 r_2 + r_1 r_3 + \cdots + r_1 r_n + r_2 r_3 + r_2 r_4 + \cdots + r_{n-1} r_n = a_2$. The pattern and its justification should now be clear and we only need a definition before this discussion can be summarized in a theorem. Let r_1, r_2, \dots, r_n be a sequence of numbers (in any field), and let k be any positive integer $1 \leq k \leq n$. Then we denote by $\sum r_1 r_2 \cdots r_k$ the sum of all the products of the form $r_{i_1} r_{i_2} \cdots r_{i_k}$ where $1 \leq i_1 < i_2 < \cdots < i_k \leq n$. Thus,

$$\begin{aligned} \sum r_i &= r_1 + r_2 + \cdots + r_n, \\ \sum r_1 r_2 &= r_1 r_2 + r_1 r_3 + \cdots + r_1 r_n + r_2 r_3 + r_2 r_4 + \cdots + r_{n-1} r_n, \\ \sum r_1 r_2 \cdots r_n &= r_1 r_2 \cdots r_n. \end{aligned}$$

At this point it is convenient to extend the notion of a polynomial to several variables. If x, y, z, \dots are variables, then any function that is obtained by adding, subtracting, or multiplying these variables and/or elements of the ground field F is called a polynomial over F . Thus,

$$197x^6y^7z^{73} + \frac{33}{17}x^2 - \frac{xy + yz + xz}{6}$$

is a polynomial over any field F in which 17, 3, and 2 are all distinct from 0.

The polynomials $\sum r_i, \sum r_1 r_2, \dots, \sum r_1 r_2 \cdots r_n$ are called the *elementary symmetric polynomials*. The above considerations prove the following theorem.

Theorem 6.19 Suppose that

$$P(x) = x^n + a_1 x^{n-1} + a_2 x^{n-2} + \cdots + a_{n-1} x + a_n = (x - r_1)(x - r_2) \cdots (x - r_n).$$

Then $\sum r_1 r_2 \cdots r_k = (-1)^k a_k$ for $k = 1, 2, \dots, n$.

This theorem can of course be used to conclude that the sum of the roots of the equation $x^3 + 6x + 5 = 0$ is 0 (the coefficient of x^2) and that their product is $(-1)^3 5 = -5$. It can, however, be brought to bear on other interesting expressions as well. Let $\sum r_i^\alpha r_j^\beta r_k^\gamma \cdots$ denote the sum of all the distinct monomials obtained by permuting the indices i, j, k, \dots in the monomial $\sum r_i^\alpha r_j^\beta r_k^\gamma \cdots$. Then $\sum r_i^\alpha r_j^\beta r_k^\gamma \cdots$ can be expressed in terms of the elementary symmetric polynomials. This general fact, which is known as the *Fundamental Theorem of Symmetric Polynomials*, will not be proved here.

Instead, some special cases will be considered. The expression $\sum r_i^2$, which denotes the sum of the squares of the zeroes of the polynomial $P(x)$ of Equation 6.16 can be evaluated, with the help of the Multinomial Theorem, as follows:

$$\sum r_i^2 = \left(\sum r_i\right)^2 - 2 \sum r_1 r_2 = a_1^2 - 2(-1)^2 a_2 = a_1^2 - 2a_2.$$

In particular, the sum of the squares of the roots of the equation $x^3 + 6x + 5 = 0$ is $0^2 - 2 \cdot 6 = -12$. Similarly, if r_1, r_2, r_3, r_4 are the roots of the equation $x^4 + 5x^3 - 3x^2 + 7x + 10 = 0$, then

$$\frac{1}{r_1} = \frac{r_1 r_2 r_3}{r_1 r_2 r_3 r_4} = -\frac{7}{10}.$$

The next proposition provides the general framework for verifying that any proposed set of roots does indeed constitute a complete solution set.

Proposition 6.20 If r_1, r_2, \dots, r_n are elements of the field F and

$$P(x) = x^n + a_1 x^{n-1} + a_2 x^{n-2} + \cdots + a_{n-1} x + a_n$$

is a polynomial over F such that

$$\sum r_1 r_2 \cdots r_k = (-1)^k a_k \quad \text{for } k = 1, 2, \dots, n,$$

then $P(x) = (x - r_1)(x - r_2) \cdots (x - r_n)$.

Proof. Set

$$Q(x) = (x - r_1)(x - r_2) \cdots (x - r_n) = x^n + b_1 x^{n-1} + b_2 x^{n-2} + \cdots + b_{n-1} x + b_n.$$

Then, by Theorem 6.19, for $k = 1, 2, \dots, n$,

$$b_k = (-1)^k \sum r_1 r_2 \cdots r_k = (-1)^k (-1)^k a_k = a_k.$$

Hence $P(x)$ and $Q(x)$ are identical polynomials. ■

We are now in position to eliminate the rough edges of the solution in Chapter 3 of the cubic equation. That solution was incomplete in that it was not proved that the three roots x_1 , x_2 , and x_3 selected in Equation 3.6 constitute the complete solution set of the general cubic equation $x^3 + ax^2 + bx + c = 0$.

Corollary 6.21 The values x_1 , x_2 , and x_3 given by Equation 3.6 are the complete solution set of the cubic equation $x^3 + ax^2 + bx + c = 0$.

Proof. Since the reduced cubic equation $y^3 + py + q = 0$ was obtained by setting $x = y - a/3$, it follows from the same Equation 3.6 that it suffices to show that $\{y_1, y_2, y_3\}$ is the complete solution set of this reduced cubic equation. By Proposition 6.20 it suffices to prove that $y_1 + y_2 + y_3 = 0$, $y_1 y_2 + y_2 y_3 + y_3 y_1 = p$, and $y_1 y_2 y_3 = -q$. We demonstrate only the last of these equalities, leaving the other two to Exercise 6.4.31. It follows from Equation 3.6 that

$$\begin{aligned} y_1 y_2 y_3 &= \left(z_1 - \frac{p}{3z_1}\right) \left(\omega z_1 - \frac{p\omega^2}{3z_1}\right) \left(\omega^2 z_1 - \frac{p\omega}{3z_1}\right) \\ &= \omega^3 z_1^3 - \frac{p z_1}{3} (\omega^3 + \omega^2 + \omega^4) + \frac{p^2}{9 z_1} (\omega^3 + \omega^2 + \omega^4) - \frac{p^3 \omega^3}{27 z_1^3} \\ &= z_1^3 - \frac{p z_1}{3} (1 + \omega^2 + \omega) + \frac{p^2}{9 z_1} (1 + \omega^2 + \omega) - \frac{p^3}{27 z_1^3} = z_1^3 - \frac{p^3}{27 z_1^3}. \end{aligned}$$

Now, by Equation 3.4, z_1^3 is one of the roots of the quadratic equation $u^2 + qu - p^3/27 = 0$ and hence the other one is $-p^3/27 z_1^3$. Since the sum of the roots of this quadratic is $-q$, it now follows that $z_1^3 - p^3/27 z_1^3 = -q$, and so $y_1 y_2 y_3 = -q$. ■

Exercises 6.4

Let r , s , and t be the roots of the equation $x^3 + ax^2 + bx + c = 0$. Rewrite the expressions in Exercises 6.4.1 to 6.4.7 in terms of a , b , and c . (Wherever necessary, you may assume that the denominators are not zero.)

1. $r^2 + s^2 + t^2$
2. $(r+s)^2 + (r+t)^2 + (s+t)^2$
3. $(r+s)(r+t)(s+t)$
4. $r^2s^2 + r^2t^2 + s^2t^2$
5. $1/r + 1/s + 1/t$
6. $1/r^2 + 1/s^2 + 1/t^2$
7. $1/(r+s) + 1/(r+t) + 1/(s+t)$

Let r , s , and t be the zeroes of the real polynomial $x^3 + 23x + 1$. Find the real cubic polynomial whose zeroes are those that appear in Exercises 6.4.8 to 6.4.13.

8. r^2, s^2, t^2
9. $r+s, s+t, t+r$
10. rs, st, tr
11. $1+r, 1+s, 1+t$
12. kr, ks, kt
13. $k/r, k/s, k/t$

Let r_1, r_2, r_3 , and r_4 be the zeroes of the polynomial $x^4 + 3x^3 - 5x + 1$. Evaluate the expressions in Exercises 6.4.14 to 6.4.16.

14. $r_1^2 + r_2^2 + r_3^2 + r_4^2$
15. $1/r_1 + 1/r_2 + 1/r_3 + 1/r_4$
16. $(r_1r_2 - r_3r_4)^2 + (r_1r_3 - r_2r_4)^2 + (r_1r_4 - r_2r_3)^2$
17. Solve the equation $x^5 - 4x^4 + 2x^3 - 8x^2 - 35x + 140 = 0$ whose solutions have the form $r, -r, s, -s, t$.
18. The equation $x^5 + 3x^4 - x^3 - 11x^2 - 17x - 5 = 0$ has two roots whose product is 1. Determine these two roots.
19. The equation $x^5 - 409x + 285 = 0$ has two roots whose sum equals 5. Determine these two roots.

Let $r_1, r_2, r_3, \dots, r_n$ be the zeroes of the polynomial $x^n + a_1x^{n-1} + a_2x^{n-2} + \dots + a_{n-1}x + a_n$. Prove the identities in Exercises 6.4.20 to 6.4.25.

20. $\sum r_1^2 r_2 = 3a_3 - a_1 a_2$
21. $\sum r_1^3 = 3a_1 a_2 - a_1^3 - 3a_3$
22. $\sum r_1^2 r_2 r_3 = a_1 a_3 - 4a_4$
23. $\sum r_1^2 r_2^2 = a_2^2 - 2a_1 a_3 + 6a_4$
24. $\sum r_1^3 r_2 = a_1^2 a_2 - 2a_2^2 - a_1 a_3 - 4a_4$
25. $\sum r_1^4 = a_1^4 - 4a_1^2 a_2 + 2a_2^2 + 4a_1 a_3 - 4a_4$

26. Let r_1, r_2, \dots, r_n be the zeroes of the polynomial $x^n + a_1 x^{n-1} + a_2 x^{n-2} + \dots + a_{n-1} x + a_n$ where $a_n \neq 0$. Prove that the zeroes of

$$x^n + \frac{a_{n-1}}{a_n} x^{n-1} + \frac{a_{n-2}}{a_n} x^{n-2} + \dots + \frac{a_1}{a_n} x + \frac{1}{a_n}$$

are $1/r_1, 1/r_2, \dots, 1/r_n$.

Let $r_1, r_2, r_3, \dots, r_n$ be the zeroes of the polynomial $x^n + a_1 x^{n-1} + a_2 x^{n-2} + \dots + a_{n-1} x + a_n$ where $a_n \neq 0$. Express the sums in Exercises 6.4.27 to 6.4.30 in terms of a_1, a_2, \dots, a_n .

27. $\sum 1/r_i$ 28. $\sum 1/r_i^2$ 29. $\sum 1/r_i r_j$ 30. $\sum 1/r_i^3$

31. Complete the proof of Corollary 6.21.

6.5 Lagrange's Solution of the Quartic Equation

In 1771 Lagrange wrote a lengthy treatise titled *Reflexions sur la Résolution Algébrique des Equations* in which he summarized what was known about solvability of equations by radicals. He also added some thoughts of his own and in fact proved several theorems that eventually did lead to the resolution of this issue by the next generation of mathematicians. It is with this contribution of Lagrange's as well as some of its subsequent developments that most of the rest of this book is concerned.

One of the methods that Lagrange offered for the solution of quartic equations began with the seemingly innocuous observation that when the roots r_1, r_2, r_3 , and r_4 are permuted (in other words, substituted for each other), the expression $r_1 r_2 + r_3 r_4$ assumes only three values, namely, itself, $r_1 r_3 + r_2 r_4$, and $r_1 r_4 + r_2 r_3$.

For example, when the variables are interchanged by cycling them to the left, $r_1 r_2 + r_3 r_4$ becomes

$$r_2 r_3 + r_4 r_1 = r_1 r_4 + r_2 r_3.$$

If, on the other hand, only r_1 and r_4 are switched, the polynomial is transformed into

$$r_4 r_2 + r_3 r_1 = r_1 r_3 + r_2 r_4.$$

This fact can be used to solve the quartic in the following manner. Let r_1, r_2, r_3 , and r_4 denote the four roots of the equation $x^4 + ax^3 + bx^2 + cx + d = 0$ and set $A = r_1 r_2 + r_3 r_4$, $B = r_1 r_3 + r_2 r_4$, and $C = r_1 r_4 + r_2 r_3$. Clearly, $A + B + C = \sum r_1 r_2 = b$.

Next,

$$\begin{aligned}
 AB + AC + BC &= \\
 (r_1 r_2 + r_3 r_4)(r_1 r_3 + r_2 r_4) &+ (r_1 r_2 + r_3 r_4)(r_1 r_4 + r_2 r_3) + (r_1 r_3 + r_2 r_4)(r_1 r_4 + r_2 r_3) \\
 &= r_1^2 r_2 r_3 = \left(\sum r_1\right)\left(\sum r_1 r_2 r_3\right) - 4\left(\sum r_1 r_2 r_3 r_4\right) = (-a)(-c) - 4d = ac - 4d.
 \end{aligned}$$

Finally,

$$\begin{aligned}
 ABC &= (r_1 r_2 + r_3 r_4)(r_1 r_3 + r_2 r_4)(r_1 r_4 + r_2 r_3) \\
 &= r_1^3 r_2 r_3 r_4 + r_1^2 r_2^2 r_3^2 \\
 &= r_1 r_2 r_3 r_4 r_1^2 + (r_1 r_2 r_3)^2 - 2r_1^2 r_2^2 r_3 r_4 \\
 &= d[(r_1)^2 - 2r_1 r_2] + (-c)^2 - 2r_1 r_2 r_3 r_4 r_1 r_2 \\
 &= d(a^2 - 2b) + c^2 - 2d b \\
 &= a^2 d + c^2 - 4b d.
 \end{aligned}$$

These computations lead to the observation that if r_1, r_2, r_3, r_4 are the roots of the quartic equation

$$x^4 + ax^3 + bx^2 + cx + d = 0,$$

then $r_1 r_2 + r_3 r_4$, $r_1 r_3 + r_2 r_4$, and $r_1 r_4 + r_2 r_3$ are the roots of the cubic equation

$$y^3 - by^2 + (ac - 4d)y - (a^2 d + c^2 - 4bd) = 0.$$

Since every cubic equation is already known to be solvable by radicals, it follows that A , B , and C have algebraic expressions in a , b , c , and d . The actual values of r_1 , r_2 , r_3 , and r_4 are now extracted with relative ease. By Exercise 6.5.8 we may assume that $r_1 r_2 \neq r_3 r_4$. Since $r_1 r_2 + r_3 r_4 = A$ and $(r_1 r_2)(r_3 r_4) = d$, it follows that $\alpha = r_1 r_2$ and $\beta = r_3 r_4$ are the solutions of the quadratic $z^2 - Az + d = 0$, and so they too have algebraic expressions in a , b , c , and d . Moreover, $(r_1 + r_2) + (r_3 + r_4) = -a$ and

$$r_3 r_4(r_1 + r_2) + r_1 r_2(r_3 + r_4) = r_1 r_2 r_3 = -c.$$

When this system of simultaneous equations is solved for $(r_1 + r_2)$ and $(r_3 + r_4)$ we get

$$r_1 + r_2 = \frac{c - ar_1r_2}{r_1r_2 - r_3r_4} = \frac{c - a\alpha}{\alpha - \beta} = \gamma$$

and

$$r_3 + r_4 = \frac{ar_3r_4 - c}{r_1r_2 - r_3r_4} = \frac{a\beta - c}{\alpha - \beta} = \delta.$$

Note that γ and δ both have algebraic expressions in a , b , c , and d . Finally, since $r_1 + r_2 = \gamma$ and $r_1r_2 = \alpha$, it follows that r_1 and r_2 are the solutions of the quadratic $u^2 - \gamma u + \alpha = 0$, and similarly r_3 and r_4 are the solutions of the quadratic $v^2 - \delta v + \beta = 0$. Thus we have proved the following theorem.

Theorem 6.22 Every fourth-degree equation is solvable by radicals.

We illustrate *Lagrange's method* with an equation whose solutions can, of course, be found in a much shorter way. Consider the equation $x^4 - x^2 = 0$. Here, the auxiliary cubic turns out to be $y^3 + y^2 = 0$ and we choose $A = -1$ (usually the choice is arbitrary, though in this case choosing $A = 0$ would lead to problems). This gives us the auxiliary quadratic $z^2 + z = 0$ and we set $\alpha = -1$ and $\beta = 0$. This gives us $\gamma = \delta = 0$ and so the roots of the original quartic are those of the quadratics $u^2 - 1 = 0$ and $v^2 = 0$, namely, $\pm 1, 0, 0$.

Exercises 6.5

Use Lagrange's method to solve the equations in Exercises 6.5.1 to 6.5.4.

1. $x^4 - 1 = 0$ 2. $x^4 + 1 = 0$ 3. $x^4 - x = 0$ 4. $x^4 + x = 0$

Explain why the equations in Exercises 6.5.5 to 6.5.7 are resolvable by radicals.

5. $x^8 - 2x^7 + 3x^6 - 5x^5 + 11x^4 - 5x^3 + 3x^2 - 2x + 1 = 0$

6. $x^8 - 3x^6 + 5x^4 - 2x^2 + 17 = 0$

7. $x^{12} - x^9 + 5x^6 - 32 = 0$

8. Suppose the roots of $x^4 + ax^3 + bx^2 + cx + d = 0$ are such that the product of any two equals the product of the other two. Prove that $x^4 + ax^3 + bx^2 + cx + d$ factors into either $(x + r)^4$ or $(x^2 - r^2)^2$ for some complex number r .

Chapter Summary

The notion of a field was isolated and identified as the underlying structure common to the rational, real, complex, and prime modulus number systems. Polynomials were studied in this abstract context, and the correspondence between their zeroes and their linear factors was pointed out. The Euclidean algorithm was demonstrated to contain much useful information about polynomials. Some attention was paid to the relation between the coefficients of a polynomial and its zeroes. It was also shown that information of this type can be used to yield a formula for the solution of the general quartic equation.

Chapter Review Exercises

Mark the following true or false.

- 1. The remainder of $x^{17} + 1$ when divided by $x^2 + 1$ is $x^3 + 1$.
- 2. In $\mathbb{R}[x]$, the remainder of $x^{72} + 1$ when divided by $x + 2$ is 4,321.
- 3. In $\mathbb{Z}_5[x]$, $x^4 - x$ is divisible by $x^2 - 3x + 2$.
- 4. $x^2 + 1$ is reducible over \mathbb{Z}_2 .
- 5. $x^2 + 1$ is reducible over every field.
- 6. The equation $5x^4 + (1 + i)x^3 - ix + 17 = 0$ has four complex roots.
- 7. The greatest common divisor of $x^2 + 1$ and $x^2 - ix$ over \mathbb{C} is $x - i$.
- 8. The product of all the complex roots of the equation $x^{17} - 3x + 5 = 0$ is -5 .
- 9. The quartic equation is solvable by radicals.

New Terms

constant polynomial, 103	ground field, 102
degree, 103	irreducible, 108
division of polynomials, 103	Lagrange's method, 127
divisors, 108	monic polynomial, 103
elementary symmetric polynomials, 121	multiplicity of a zero, 110
Euclidean greatest common divisor, 115	polynomial over a field, 101
factorable, 108	reducible, 108
factorization, 108	relatively prime, 116
factors, 108	variables, 101
field, 99	zero of a polynomial, 107
greatest common divisor, 113	zero polynomial, 103

Supplementary Exercises

1. Write a computer script that finds the greatest common divisor of any two polynomials with real coefficients.
2. Write a computer script that finds the greatest common divisor of any two polynomials with coefficients in \mathbb{Z}_2 .
3. Write a computer script that finds the greatest common divisor of any two polynomials with coefficients in \mathbb{Z}_p .
4. Write a computer script that implements Lagrange's solution of the general quartic with complex coefficients.
5. Write a computer script that lists the monic irreducible polynomials of degree d over \mathbb{Z}_p .
6. Find a formula for the number irreducible monic polynomials of degree d over \mathbb{Z}_p .
7. Prove that every symmetric polynomial in the variables x_1, x_2, \dots, x_n is expressible as a polynomial in the elementary symmetric polynomials

$$x_1, x_1x_2, \dots, x_1x_2 \cdots x_n.$$

Chapter 7



GALOIS FIELDS

SOME NEW FIELDS are introduced and studied in detail. These fields combine some of the features of both the complex and the modular numbers systems. The existence of primitive roots modulo p is proved in this new context.

7.1 Galois's Construction of His Fields

The following quotation consists of the opening paragraphs of the article *On the Theory of Numbers* by Évariste Galois, which appeared in the June 1830 issue of the *Bulletin des Sciences mathématiques*. Some of the notation has been modernized for pedagogical reasons and a more faithful translation appears in Appendix D.

When it is agreed to consider as zero all the quantities which are the multiples of a given prime number p , and, subject to this convention, one looks for solutions to the polynomial equation $F(x) = 0$, i.e., the equations that Mr. Gauss denotes by $F(x) \equiv 0$, it is customary to consider only integer solutions to these sorts of questions. Having been led by some specific researches to consider their irrational solutions, I have arrived at some results that I consider to be new.

Let there be given such an equation or congruence, $F(x) = 0$, and let p be the modulus. Suppose first that the congruence in question admits no rational factors, that is, there exist no three polynomials $\varphi(x)$, $\psi(x)$, $\chi(x)$ such that

$$\varphi(x) \cdot \psi(x) = F(x) + p \cdot \chi(x).$$

In that case the congruence has no integer roots, nor any factor of smaller degree. One should therefore regard the roots of this congruence as some kind of imaginary symbols (since they do not satisfy the same questions as integers), symbols whose employment, in calculations, will often prove as useful as that of the imaginary $\sqrt{-1}$ in ordinary analysis. We are concerned here with the classification of these imaginaries and the minimization of their number. Let i denote one of the roots of the congruence $F(x) = 0$, which can be supposed to have degree ν .

Consider the general expression

$$a_0 + a_1 i + a_2 i^2 + \cdots + a_{\nu-1} i^{\nu-1}, \quad (\text{A})$$

where $a_0, a_1, a_2, \dots, a_{\nu-1}$ represent integers. When these numbers are assigned all their possible values, Expression A runs through p^ν values which possess, as I shall demonstrate, the same properties as the natural numbers in the theory of residues of powers.

In the first paragraph Galois states that it is his intention to consider the solutions of polynomial equations with coefficients in \mathbb{Z}_p . He then goes on to explain what is meant by irreducibility of polynomials modulo p , a topic that was covered in Section 6.2. The closing sentence of the second paragraph is extraordinarily creative and imaginative. Just as the polynomial $x^2 + 1$, which is irreducible over the real numbers, yields the imaginary but useful number $\sqrt{-1}$, so do these polynomials which are irreducible over \mathbb{Z}_p yield a new species of imaginary symbols. Accordingly, we shall refer to these as *Galois imaginaries*.

Galois next proceeds to draw some further consequences from this analogy. It was seen in Section 2.1 that if $\sqrt{-1}$ is a zero of the irreducible quadratic $x^2 + 1$, i.e., if

$$(\sqrt{-1})^2 + 1 = 0 \quad \text{or} \quad (\sqrt{-1})^2 = -1, \quad (7.1)$$

then any rational function of $\sqrt{-1}$ can be reduced to the form $a + b\sqrt{-1}$. Galois now asserts that if i is the new imaginary number associated with an irreducible polynomial $F(x)$ of degree ν over \mathbb{Z}_p , then every rational function of i can be reduced to the form

$$a_0 + a_1 i + a_2 i^2 + \cdots + a_{\nu-1} i^{\nu-1}. \quad (7.2)$$

Let us digress here into some concrete computations. Consider the irreducible polynomial $x^2 + x + 1$ over \mathbb{Z}_2 , and suppose it has α as a Galois imaginary, so that, in analogy with Equation 7.1, α satisfies the equation $\alpha^2 + \alpha + 1 = 0$ over \mathbb{Z}_2 , or $\alpha^2 = 1 + \alpha$. Consequently, any second-degree polynomial in α can be reduced to a linear function of α . For example,

$$\alpha^2 + 1 = 1 + \alpha + 1 = 2 + \alpha = 0 + \alpha = \alpha.$$

The same holds for cubic polynomials, since

$$\alpha^3 = \alpha^2 \alpha = (1 + \alpha) \alpha = \alpha + \alpha^2 = \alpha + 1 + \alpha = 1 + 2\alpha = 1 + 0\alpha = 1.$$

Similarly, $\alpha^4 = \alpha^3\alpha = 1\alpha = \alpha$, $\alpha^5 = \alpha^4\alpha = \alpha\alpha = \alpha^2 = 1 + \alpha$, and so on. In other words, the successive powers of α cycle through the values 1, α , and $1 + \alpha$, and hence every polynomial function of α can be reduced to the form

$$\alpha_0 + \alpha_1\alpha \quad \text{with } \alpha_0, \alpha_1 \in \mathbb{Z}_2. \quad (7.3)$$

The preceding considerations make it clear that the sum and product of any two of the expressions 0, 1, α , $\alpha + 1$ is again an expression of the same form. Let us now examine the issue of division. Does each of the nonzero elements of the form in Expression 7.3 have an inverse? Since the coefficients α_0 and α_1 can assume only the values 0 or 1, there are only three nonzero elements to consider: $1 = \alpha^0$, $\alpha = \alpha^1$, and $1 + \alpha = \alpha^2$. As it is already known that $\alpha^3 = 1$, it follows that $1^{-1} = 1$ (of course), $\alpha^{-1} = \alpha^2$, and $(\alpha^2)^{-1} = \alpha^1 = \alpha$. Thus the elements 0, 1, α , and $1 + \alpha$ form a field provided that 0 and 1 are understood to be elements of \mathbb{Z}_2 , and provided it is assumed that $\alpha^2 + \alpha + 1 = 0$. With the sole exception of the issue of the existence of multiplicative inverses, the validity of the field properties in this new context follows from their validity for polynomials, since $1 + \alpha$ is treated much the same as the polynomial $1 + x$, etc.

Let us consider another example in detail before we comment on Galois's brainchild in general. The polynomial $x^3 + x^2 + 1$ is irreducible of degree 3 over \mathbb{Z}_2 . Let β be a Galois imaginary associated with this polynomial. I.e., β is a number such that $\beta^3 + \beta^2 + 1 = 0$, or $\beta^3 = 1 + \beta^2$, over \mathbb{Z}_2 . Then

$$\beta^4 = \beta^3\beta = (1 + \beta^2)\beta = \beta + \beta^3 = \beta + 1 + \beta^2 = 1 + \beta + \beta^2,$$

$$\beta^5 = \beta^4\beta = (1 + \beta + \beta^2)\beta = \beta + \beta^2 + \beta^3 = \beta + \beta^2 + 1 + \beta^2 = 1 + \beta + 2\beta^2 = 1 + \beta,$$

$$\beta^6 = \beta^5\beta = (1 + \beta)\beta = \beta + \beta^2,$$

$$\beta^7 = \beta^6\beta = (\beta + \beta^2)\beta = \beta^2 + \beta^3 = \beta^2 + 1 + \beta^2 = 1 + 2\beta^2 = 1.$$

It is again clear that every polynomial function of β is reducible to the form

$$\alpha_0 + \alpha_1\beta + \alpha_2\beta^2,$$

where each of the coefficients α_0 , α_1 , and α_2 can assume the values 0 or 1. There are exactly $2^3 = 8$ such values, and, as seen above, these eight values can also be listed as

$0, 1, \beta, \beta^2, \beta^3, \beta^4, \beta^5, \beta^6$. Since the fact that $\beta^7 = 1$ implies that

$$(\beta^k)^{-1} = \beta^{7-k} \quad \text{for all } k = 0, 1, 2, 3, 4, 5, 6,$$

it follows that this set of eight Galois imaginaries constitutes a field.

It turns out that the set of elements of the form of Expression 7.2 generated by a Galois imaginary is always a field, a fact that will be proved shortly. That such a set is closed with respect to addition, subtraction, and multiplication is quite clear. The existence of multiplicative inverses, however, is another, less obvious, matter. The polynomial $x^4 + x^3 + x^2 + x + 1$ is irreducible over \mathbb{Z}_2 and so it has a Galois imaginary γ associated with it, such that $\gamma^4 = 1 + \gamma + \gamma^2 + \gamma^3$ over \mathbb{Z}_2 . The form of Expression 7.2 gives rise to $2^4 = 16$ associated elements. However, if we now proceed to list the powers of γ , as was done earlier for α and β , we encounter a difficulty, for

$$\gamma^5 = \gamma^4 \gamma = (1 + \gamma + \gamma^2 + \gamma^3) \gamma = \gamma + \gamma^2 + \gamma^3 + \gamma^4 = \gamma + \gamma^2 + \gamma^3 + 1 + \gamma + \gamma^2 + \gamma^3 = 1.$$

For the first time the successive powers of the Galois imaginary have failed to cycle through all the elements of the form of Expression 7.2. Consequently, the device used in the previous examples to identify the inverse of each nonzero element is not available here. Nevertheless, the Galois imaginary γ does generate a corresponding field. To demonstrate this it is only necessary to show that when the symbol i of Expression 7.2 is replaced by γ , or by any Galois imaginary, then each nonzero element of the form of Expression 7.2 does indeed have a multiplicative inverse of the same form. The same Euclidean algorithm that was used to prove the existence of inverses in \mathbb{Z}_p works in this new context as well.

Lemma 7.4 Let $P(x)$ be an irreducible polynomial of degree ν over \mathbb{Z}_p , and let δ be the associated Galois imaginary. For each element

$$\zeta = a_0 + a_1 \delta + a_2 \delta^2 + \cdots + a_{\nu-1} \delta^{\nu-1}, \quad a_i \in \mathbb{Z}_p, i = 0, 1, 2, 3, \dots,$$

if the coefficients $a_0, a_1, \dots, a_{\nu-1}$ are not all zero, then there exists an element η of the same form such that $\eta \zeta = 1$.

Proof. With ζ as above, define

$$Q(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_{\nu-1} x^{\nu-1}.$$

Since $P(x)$ is an irreducible polynomial of degree ν and $Q(x)$ is a nonzero polynomial of degree less than ν , it follows that $P(x)$ and $Q(x)$ have 1 as a greatest common divisor. Hence, by Proposition 6.15 there exist polynomials $A(x)$ and $B(x)$ such that

$$A(x)Q(x) + B(x)P(x) = 1 \quad \text{over } \mathbb{Z}_p.$$

Consequently, $A(\delta)Q(\delta) + B(\delta)P(\delta) = 1$ over \mathbb{Z}_p . However, by the definition of δ , $P(\delta) = 0$, and by the definition of $Q(x)$, $Q(\delta) = \zeta$. Hence, $A(\delta)\zeta = 1$ over \mathbb{Z}_p , and so we can choose $\eta = A(\delta)$. ■

If $P(x)$ is any irreducible polynomial over \mathbb{Z}_p and i is a corresponding Galois imaginary, then the set of all the elements of the form of Expression 7.2 is denoted by $\text{GF}(p, P(x))$ and is called a *Galois field*. Thus, the three Galois fields described above are $\text{GF}(2, x^2 + x + 1)$, $\text{GF}(2, x^3 + x^2 + x + 1)$, and $\text{GF}(2, x^4 + x^3 + x^2 + x + 1)$. It follows from Lemma 7.4 that every Galois field is indeed a field in the sense of Section 6.1, and we state this explicitly.

Theorem 7.5 If $P(x)$ is an irreducible polynomial over \mathbb{Z}_p , then the Galois field $\text{GF}(p, P(x))$ is a field.

Just as the polynomial $x^2 + 1$, which is irreducible over \mathbb{R} , has the two imaginaries i and $-i$ associated with it, so is it the case that every degree ν polynomial $P(x)$ that is irreducible over \mathbb{Z}_p has ν distinct Galois imaginaries. However, it so happens that the Galois field generated by any one of these imaginaries of $P(x)$ contains all the other $\nu - 1$ imaginaries of $P(x)$, and so all of the Galois imaginaries of $P(x)$ define the same Galois field $\text{GF}(p, P(x))$. A more detailed discussion of this phenomenon will be found in Section 7.4 below.

Galois does not prove Theorem 7.5 explicitly. Later in his paper he writes

Next it can be proven, just as is done in the theory of numbers, that there exist primitive roots $\alpha \dots$ which \dots reproduce, by their powers, the complete sequence of all the other roots.

In other words, Galois claims that the Galois field $\text{GF}(p, P(x))$ contains an element α whose powers $1, \alpha, \alpha^2, \alpha^3, \dots$ run through all the nonzero values of Expression 7.2 above. Because of the similarity that this bears to the powers of the complex roots of unity, such an element is also called *primitive*. The examples preceding Lemma 7.4 show how such a primitive element can be used to demonstrate the existence of multiplicative inverses. A table that expresses all the powers of a primitive Galois imaginary in the form of Expression 7.2 is called a *cyclic table*. Table 7.1 contains the cyclic table of the primitive

$\beta^0 = 1$	$\beta^4 = 1 + \beta + \beta^2$
$\beta^1 = \beta$	$\beta^5 = 1 + \beta$
$\beta^2 = \beta^2$	$\beta^6 = \beta + \beta^2$
$\beta^3 = 1 + \beta^2$	$\beta^7 = 1$

Table 7.1 The cyclic table of $\text{GF}(2, x^3 + x^2 + 1)$

Galois imaginary β associated with $\text{GF}(2, x^3 + x^2 + 1)$. The detailed calculations were displayed above.

The elements α and β of this chapter's first two examples are primitive elements of $\text{GF}(2, x^2 + x + 1)$ and $\text{GF}(2, x^3 + x^2 + 1)$ respectively. On the other hand, the element γ fails to be a primitive element of $\text{GF}(2, x^4 + x^3 + x^2 + x + 1)$. The *order* $\text{o}(\zeta)$ of any element ζ of some Galois field is the least positive integer k such that $\zeta^k = 1$. Thus, in the aforementioned Galois field γ has order 5. On the other hand, in $\text{GF}(2, x^2 + x + 1)$, the elements α and $\alpha + 1$ both have order 3. Similarly, with the exception of 0 and 1, each of the elements of $\text{GF}(2, x^3 + x^2 + 1)$ has order 7. As it would be pedagogically useful to have at least one Galois field that is not associated with a polynomial over \mathbb{Z}_2 , we compute the cyclic table of $\text{GF}(3, x^2 + x + 2)$. That this polynomial is indeed irreducible over \mathbb{Z}_3 follows from the facts that

$$0^2 + 0 + 2 = 2 \neq 0,$$

$$1^2 + 1 + 2 = 4 \neq 0$$

and

$$2^2 + 2 + 2 = 8 \not\equiv 0 \pmod{3}.$$

If σ is the associated Galois imaginary, then

$$\sigma^2 + \sigma + 2 = 0$$

over \mathbb{Z}_3 , and so

$$\sigma^2 = -2 - \sigma = 1 + 2\sigma.$$

Consequently

$$\sigma^3 = \sigma^2 \sigma = (1 + 2\sigma)\sigma = \sigma + 2\sigma^2 = \sigma + 2(1 + 2\sigma) = 2 + 5\sigma = 2 + 2\sigma,$$

$$\sigma^4 = \sigma^3 \sigma = (2 + 2\sigma)\sigma = 2\sigma + 2\sigma^2 = 2\sigma + 2(1 + 2\sigma) = 2 + 6\sigma = 2,$$

$$\sigma^5 = \sigma^4 \sigma = 2\sigma,$$

$$\sigma^6 = \sigma^5 \sigma = 2\sigma^2 = 2 + 4\sigma = 2 + \sigma,$$

$$\sigma^7 = \sigma^6 \sigma = (2 + \sigma)\sigma = 2\sigma + \sigma^2 = 2\sigma + 1 + 2\sigma = 1 + \sigma,$$

$$\sigma^8 = \sigma^7 \sigma = (1 + \sigma)\sigma = \sigma + \sigma^2 = \sigma + 1 + 2\sigma = 1.$$

It follows that σ is indeed a primitive element of $\text{GF}(3, x^2 + x + 2)$.

The following proposition provides us with some general information about the number of elements of a Galois field and their orders.

Proposition 7.6 Let F be a Galois field associated with the irreducible polynomial $P(x)$ of degree ν over \mathbb{Z}_p . Then F has exactly p^ν elements and the order of each nonzero element of F is finite.

Proof. Let i be the Galois imaginary associated with $P(x)$. As was remarked above, every element of F is expressible in the form of Expression 7.2. Hence F contains at most p^ν elements. In order to show that F contains exactly that number of elements, it suffices to show that no two distinct expressions of the form of Expression 7.2 can equal each other. Suppose ζ_1 and ζ_2 are distinct elements of F . Then

$$\zeta_1 = a_0 + a_1 i + a_2 i^2 + \cdots + a_{\nu-1} i^{\nu-1}$$

and

$$\zeta_2 = b_0 + b_1 i + b_2 i^2 + \cdots + b_{\nu-1} i^{\nu-1}$$

where $a_k \neq b_k$ for some $k = 0, 1, \dots, \nu - 1$. Then, when Lemma 7.4 is applied to the element

$$\zeta = \zeta_1 - \zeta_2 = (a_0 - b_0) + (a_1 - b_1)i + (a_2 - b_2)i^2 + \cdots + (a_{\nu-1} - b_{\nu-1})i^{\nu-1},$$

we conclude that there is an element η such that $\zeta\eta = 1$. By Proposition 6.1, ζ is not zero and so $\zeta_1 \neq \zeta_2$. Thus F has exactly p^ν elements.

Let α be any nonzero element of F . Since all the terms of the infinite sequence $1, \alpha, \alpha^2, \alpha^3, \dots$ are elements of F , and since F contains only a finite number of elements,

it follows that there exist two distinct exponents $m > n$ such that $\alpha^m = \alpha^n$. But then $\alpha^{m-n} = 1$, and so $\text{o}(\alpha) \leq m - n$. Thus the order of each nonzero element of F is finite. ■

Exercises 7.1

1. Write out the cyclic table of $\text{GF}(2, x^3 + x + 1)$.
2. Write out the cyclic table of $\text{GF}(2, x^4 + x + 1)$.

Verify that the Galois imaginary associated with each of the polynomials in Exercises 7.1.3 and 7.1.4 is primitive over \mathbb{Z}_2 and write out the associated cyclic table.

3. $x^5 + x^2 + 1$
4. $x^5 + x^3 + 1$

Verify that the Galois imaginary associated with each of the polynomials in Exercises 7.1.5 to 7.1.7 is primitive over \mathbb{Z}_3 and write out the associated cyclic table.

5. $x^2 + 2x + 2$
6. $x^3 + 2x + 1$
7. $x^3 + 2x^2 + 1$

8. Verify that the Galois imaginary associated with $x^2 + 4x + 2$ is primitive over \mathbb{Z}_5 and construct its cyclic table.
9. Verify that the Galois imaginary associated with $x^2 + 6x + 3$ is primitive over \mathbb{Z}_7 and construct its cyclic table.

Let β be the Galois imaginary associated with the irreducible polynomial $x^3 + x^2 + 1$ over \mathbb{Z}_2 . Solve the (systems of simultaneous) equations in Exercises 7.1.10 to 7.1.13 in $\text{GF}(2, x^3 + x^2 + 1)$.

10. $(1 + \beta)x + \beta = 1 + \beta^2$
11. $x + y = \beta$ and $x + \beta y = 1$
12. $x + (1 + \beta)y = \beta + \beta^2$ and $(1 + \beta^2)x + y = 0$
13. $x + \beta y + \beta^2 z = 1 + \beta$; $(1 + \beta)x + (1 + \beta^2)y + z = 1$; and $\beta y + z = \beta$
14. Solve Exercise 7.1.10 with $\text{GF}(2, x^3 + x^2 + 1)$ replaced by $\text{GF}(3, x^2 + 2x + 2)$.
15. Solve Exercise 7.1.11 with $\text{GF}(2, x^3 + x^2 + 1)$ replaced by $\text{GF}(3, x^2 + 2x + 2)$.
16. Solve Exercise 7.1.12 with $\text{GF}(2, x^3 + x^2 + 1)$ replaced by $\text{GF}(3, x^2 + 2x + 2)$.
17. Solve Exercise 7.1.13 with $\text{GF}(2, x^3 + x^2 + 1)$ replaced by $\text{GF}(3, x^2 + 2x + 2)$.
18. Explain why $\text{GF}(p, x - 1) = \mathbb{Z}_p$.
19. Explain why the Binomial Theorem holds for elements of Galois fields.
20. Prove that $(a \pm b)^p = a^p \pm b^p$ for any $a, b \in \text{GF}(p, P(x))$.

21. Prove that $\text{GF}(p, P(x))$ contains exactly one p -th root of unity.
22. Prove that for any $a, b \in \text{GF}(p, P(x))$, $a^p = b^p$ if and only if $a = b$.
23. Prove that for any positive integer k and for any $a, b \in \text{GF}(p, P(x))$, $a^{p^k} = b^{p^k}$ if and only if $a = b$.
24. True or false: for any $a, b \in \text{GF}(p, P(x))$, $a^2 = b^2$ if and only if $a = b$. Justify your answer.
25. Show that for any $a \in \text{GF}(p, P(x))$, $a^p = a$ if and only if $a \in \mathbb{Z}_p$.
26. Let k be any fixed positive integer. Show that $\{a \in \text{GF}(p, P(x)) \mid a^{p^k} = a\}$ is also a field.
27. Prove that if $p \neq 2$, then the sum of all the elements of $\text{GF}(p, P(x))$ is 0.
28. Prove that the product of all the nonzero elements of $\text{GF}(p, P(x))$ is -1 .
29. Let a be an element of the Galois field $\text{GF}(p, P(x))$. Prove that
 - (a) $1 + a + a^2 + \cdots + a^{o(a)-1} = 0$.
 - (b) $1 \cdot a \cdot a^2 \cdots a^{o(a)-1} = (-1)^{o(a)-1}$.

7.2 The Galois Polynomial

The orders of the elements of Galois fields, defined in the previous section, possess the same properties as the orders of the complex and modular roots of unity, which are restated here for the sake of completeness. Since the proofs of Proposition 2.16, Corollary 2.17, and Propositions 5.19 and 5.20 work in the new context verbatim, these properties are restated without proof.

Proposition 7.7 Let α and β be any roots of unity in some field F . Then

- (a) $\alpha^n = 1$ if and only if n is a multiple of $o(\alpha)$;
- (b) $\alpha^a = \alpha^b$ if and only if $o(\alpha)$ is a divisor of $a - b$ and so $1, \alpha, \alpha^2, \dots, \alpha^{o(\alpha)-1}$ are all distinct;
- (c) if $o(\alpha) = n$, then $o(\alpha^k) = n/(k, n)$;
- (d) $o(\alpha\beta) = o(\alpha)o(\beta)$ if $o(\alpha)$ and $o(\beta)$ are relatively prime.

If α is any element of order k , it must clearly be a zero of the polynomial $x^m - 1$ whenever m is a multiple of k . Hence, by the first part of the above proposition, if e is the least common multiple of the orders of all the nonzero elements of the Galois field $\text{GF}(p, P(x))$, then these elements are all zeroes of $x^e - 1$. This number e is, of course, of interest, and it will eventually be demonstrated (Theorem 7.17) that $e = p^\nu - 1$, where ν

is the degree of $P(x)$. We begin this process by picking up where the previous section's quotation from Galois's paper left off.

Of the expressions [in Expression A] we shall only take the $p^\nu - 1$ values obtained when $a_0, a_1, a_2, \dots, a_{\nu-1}$ are not all zero; let α be one of these expressions.

If α is successively raised to the second, third, ... powers, a sequence of quantities all of which have the same form is obtained (since every function of i is reducible to the $(\nu - 1)$ -th degree). Hence it must be that $\alpha^n = 1$ for some n ; let n be the smallest number such that $\alpha^n = 1$. Then the numbers $1, \alpha, \alpha^2, \alpha^3, \dots, \alpha^{n-1}$ are all distinct. Next, multiply these n numbers by another expression C of the same form. We then obtain another new group of quantities all different from the first group as well as from each other. If the quantities of Form 7.2 have not been exhausted yet, the powers of α can be multiplied by a new expression γ , and so on. Consequently the number n necessarily divides the total number of quantities of Form 7.2. Since this number is $p^\nu - 1$, we see that n divides $p^\nu - 1$. From this it also follows that

$$\alpha^{p^\nu - 1} = 1, \quad \text{or} \quad \alpha^{p^\nu} = \alpha.$$

Two sentences later we find the following statement:

We note here the remarkable result that all the algebraic quantities that arise in this theory are roots of equations of the form

$$x^{p^\nu} = x.$$

To illustrate the above procedure, consider the Galois imaginary γ associated with the polynomial $x^4 + x^3 + x^2 + x + 1$ which is irreducible over \mathbb{Z}_2 . We saw that $\gamma^5 = 1$. Since $\gamma^4 = 1 + \gamma + \gamma^2 + \gamma^3$, the process begins with

$$1, \gamma, \gamma^2, \gamma^3, 1 + \gamma + \gamma^2 + \gamma^3. \quad (7.8)$$

According to Proposition 7.6, any two elements of Expression A are distinct, and hence $1 + \gamma$ is different from all the elements listed in Equation 7.8. Using this $1 + \gamma$ as C , the next set of elements produced by Galois's procedure is

$$(1 + \gamma)1, (1 + \gamma)\gamma, (1 + \gamma)\gamma^2, (1 + \gamma)\gamma^3, (1 + \gamma)(1 + \gamma + \gamma^2 + \gamma^3),$$

or, upon simplification,

$$1 + \gamma, \gamma + \gamma^2, \gamma^2 + \gamma^3, 1 + \gamma + \gamma^2, \gamma + \gamma^2 + \gamma^3. \quad (7.9)$$

The element $1 + \gamma^2$ has not been listed yet, so the next list is

$$1 + \gamma^2, (1 + \gamma^2)\gamma, (1 + \gamma^2)\gamma^2, (1 + \gamma^2)\gamma^3, (1 + \gamma^2)(1 + \gamma + \gamma^2 + \gamma^3),$$

or, upon simplification,

$$1 + \gamma^2, \gamma + \gamma^3, 1 + \gamma + \gamma^3, 1 + \gamma^3, 1 + \gamma^2 + \gamma^3. \quad (7.10)$$

It is easily verified by inspection that, as Galois claims, the three sets listed in Equations 7.8, 7.9, and 7.10 exhaust all the nonzero elements of $\text{GF}(2, x^4 + x^3 + x^2 + x + 1)$. We now state Galois's theorem and supply some of the missing details in his proof. Another, much more general and succinct proof will be provided later by Proposition 9.16 and Exercise 9.5.12.

Theorem 7.11 (Galois) Let $P(x)$ be an irreducible polynomial of degree ν over \mathbb{Z}_p . Then all the elements of $\text{GF}(p, P(x))$ are zeroes of the polynomial $x^{p^\nu} - x$.

Proof. Let α be any nonzero element of $F = \text{GF}(p, P(x))$ and suppose that it has order n . By Proposition 7.7, the list

$$1, \alpha, \alpha^2, \dots, \alpha^{n-1} \quad (7.12)$$

consists of n distinct elements. If this list does not exhaust all the nonzero elements of F , let β be a nonzero element that does not appear in this list, and consider the new list

$$\beta, \alpha\beta, \alpha^2\beta, \dots, \alpha^{n-1}\beta. \quad (7.13)$$

All the elements of List 7.13 are distinct from each other since otherwise some two elements of List 7.12 would also be nondistinct. Moreover, List 7.12 and List 7.13 are also disjoint since otherwise we would have, for some integers k and m , $\alpha^k\beta = \alpha^m$, implying that β is a power of α , which we know not to be the case. If List 7.12 and List 7.13 do not exhaust the field, we choose an element γ that is in neither list and repeat the process. Since the field F is known to be finite, this process must eventually terminate and leave the nonzero elements of the field partitioned into disjoint lists each of which contains exactly n elements. It follows that n is a divisor of the total number of nonzero elements of F , which, by Proposition 7.6, is $p^\nu - 1$. Hence, by Proposition 7.7,

$$\alpha^{p^\nu - 1} = 1. \quad (7.14)$$

Thus α , an arbitrary nonzero element of F , is a zero of the polynomial $x^{p^\nu-1} - 1$. Consequently, each of the elements of F , 0 included, is a zero of the polynomial $x(x^{p^\nu-1} - 1) = x^{p^\nu} - x$. ■

As Galois notes, this is a remarkable fact, for it provides us with a single, very simple polynomial, $x^{p^\nu} - x$, that contains all the elements of $\text{GF}(p, P(x))$ as its zeroes. We shall refer to this polynomial as the *Galois polynomial* of $\text{GF}(p, P(x))$. The observation that the order of each nonzero element of F divides $p^\nu - 1$ implies that the least common multiple of all the orders, denoted by e in this section's opening paragraph, divides $p^\nu - 1$.

It should be of interest to examine the case $\nu = 1$. The irreducible polynomials over \mathbb{Z}_p of degree $\nu = 1$ are of course the binomials $x - a$ where $a \in \mathbb{Z}_p$. Now the Galois imaginary associated with $x - a$ is none other than the known quantity a , so that

$$\text{GF}(p, x - a) = \mathbb{Z}_p.$$

Here the Galois polynomial is simply $x^p - x$ and Theorem 7.11 reduces to Fermat's Theorem 5.15.

The correspondence between the linear factors of a polynomial and its zeroes yields the following result.

Corollary 7.15 If $P(x)$ is an irreducible polynomial of degree ν over \mathbb{Z}_p , then

$$x^{p^\nu} - x = \prod_{i=1}^{p^\nu} (x - \alpha_i)$$

where $\alpha_1, \alpha_2, \dots, \alpha_{p^\nu}$ is any listing of the elements of $\text{GF}(p, P(x))$.

Proof. By Theorem 7.11 the polynomial $x^{p^\nu} - x$ has the p^ν distinct elements of the field $\text{GF}(p, P(x))$ as its zeroes. By Proposition 6.8, the polynomial $x^{p^\nu} - x$ cannot have any more zeroes. Thus, $\alpha_1, \alpha_2, \dots, \alpha_{p^\nu}$ constitute all the zeroes of the Galois polynomial $x^{p^\nu} - x$. The statement of the corollary now follows from Corollary 6.7. ■

If α is the Galois imaginary of the polynomial $x^2 + x + 1$ over \mathbb{Z}_2 , then the elements of $\text{GF}(2, x^2 + x + 1)$ are 0, 1, α , $1 + \alpha$ and

$$\begin{aligned}(x-0)(x-1)(x-\alpha)(x-(1+\alpha)) &= [x(x+1)][(x+\alpha)(x+1+\alpha)] \\ &= (x^2+x)(x^2+x+\alpha^2+\alpha) = (x^2+x)(x^2+x+1) \\ &= x^4 + x^3 + x^2x^2 + x^3 + x^2 + x = x^4 + x = x^{2^2} - x.\end{aligned}$$

Exercises 7.2

1. Find the orders of all the nonzero elements of $\text{GF}(2, x^3 + x^2 + 1)$.
2. Find the orders of all the nonzero elements of $\text{GF}(2, x^4 + x + 1)$.
3. Find the orders of all the nonzero elements of $\text{GF}(2, x^5 + x^2 + 1)$.
4. Find the orders of all the nonzero elements of $\text{GF}(3, x^2 + 2x + 2)$.
5. Find the orders of all the nonzero elements of $\text{GF}(5, x^2 + 4x + 2)$.
6. Find the orders of all the nonzero elements of $\text{GF}(7, x^2 + 6x + 3)$.
7. Verify Corollary 7.15 directly for $\text{GF}(2, x^3 + x^2 + 1)$.
8. Verify Corollary 7.15 directly for $\text{GF}(3, x^2 + 2x + 2)$.
9. Show that if $\text{GF}(p, P(x))$ is a Galois field and α is any of its elements, then $\alpha^{1+p+p^2+\dots+p^{v-1}} \in \mathbb{Z}_p$, where v is the degree of $P(x)$.

Let F be the Galois field $\text{GF}(p, P(x))$ for Exercises 7.2.10 to 7.2.13.

10. Show that the sum of all the elements of F is either 0 or 1.
11. Use the Galois polynomial to prove that the product of all the nonzero elements of F is -1 .
12. What is the sum of the squares of the elements of F ?
13. Evaluate the sum of the reciprocals of all the nonzero elements of F .
14. Let p be a prime number and let n be relatively prime to $p^v - 1$. Prove that there is exactly one n -th root of unity in $\text{GF}(p, P(x))$.
15. Suppose $a, b \in \text{GF}(2, P(x))$ for some degree 9 polynomial $P(x)$ that is irreducible over \mathbb{Z}_2 . Suppose further that $a^2 + ab + b^2 = 0$. Prove that $a = b$.

7.3 The Primitive Element Theorem

Toward the end of his paper Galois lets on that his purpose in constructing these new number systems was to find new contexts within which primitive roots exist and to which Gauss's techniques, which proved so effective for the algebraic resolution of the cyclotomic equation (see Section 2.4), could be applied to produce new algebraically resolvable equations. Galois does not prove the existence of these primitive elements, contenting himself with a comment to the effect that Gauss's proof of the existence of primitive roots modulo p carries over intact to this new setting. We will not follow Gauss's proof here and give instead a more modern, and somewhat shorter, proof.

Lemma 7.16 If F is a Galois field with f elements, and if q^m is the largest power of the prime number q that divides $f - 1$, then F contains an element a of order q^m .

Proof. The polynomial $x^{(f-1)/q} - 1$ has degree $(f-1)/q < f-1$, and so it follows from Proposition 6.8 that there is a nonzero element $b \in F$ which is not a zero of this polynomial, i.e., $b^{(f-1)/q} \neq 1$. Set $a = b^{(f-1)/q^m}$. Then,

$$a^{q^{m-1}} = b^{(f-1)/q} \neq 1$$

while, by Proposition 7.6 and Theorem 7.11,

$$a^{q^m} = b^{(f-1)} = 1.$$

Thus, $\text{o}(a)$ divides q^m but not q^{m-1} , whence $\text{o}(a) = q^m$. ■

We are ready for this chapter's main theorem:

Theorem 7.17 (The Primitive Element Theorem—Galois) Every Galois field has a primitive element.

Proof. Let F be a Galois field and suppose it contains f elements. If the prime factorization of $f - 1$ is

$$f - 1 = p_1^{m_1} p_2^{m_2} \cdots p_k^{m_k},$$

then, by the lemma, there exist elements $a_1, a_2, \dots, a_k \in F$ such that $\text{o}(a_i) = p_i^{m_i}$ for all $i = 1, 2, \dots, k$. It follows from Proposition 7.7 that

$$\text{o}(a_1 a_2 \cdots a_k) = p_1^{m_1} p_2^{m_2} \cdots p_k^{m_k} = f - 1.$$

Hence $a_1 a_2 \cdots a_k$ is the required primitive element of F . ■

It was pointed out above that $\mathbb{Z}_p = \text{GF}(p, x - 1)$ so that \mathbb{Z}_p is also a Galois field and hence Theorem 7.17 guarantees the existence of primitive elements in \mathbb{Z}_p . Thus, 3 is such a primitive element of \mathbb{Z}_{17} , since its first 16 powers are

$$1, 3, 9, 10, 13, 5, 15, 11, 16, 14, 8, 7, 4, 12, 2, 6,$$

i.e., all the nonzero elements of \mathbb{Z}_{17} . This sequence is, of course, identical with the exponents in the sum

$$\zeta + \zeta^3 + \zeta^9 + \zeta^{10} + \zeta^{13} + \zeta^5 + \zeta^{15} + \zeta^{11} + \zeta^{16} + \zeta^{14} + \zeta^8 + \zeta^7 + \zeta^4 + \zeta^{12} + \zeta^2 + \zeta^6$$

that was used in Section 2.4 to prove the constructibility of the regular 17-sided polygon.

The primitive elements of \mathbb{Z}_p , whose existence is guaranteed by Theorem 7.17, are identical with the primitive roots (modulo p) that were defined in Section 5.2. It was Euler who first proved the existence of these primitive roots modulo p . Gauss expanded on this work of Euler's and also applied it to his analysis of the cyclotomic equation $x^p - 1 = 0$. Galois, in attempting to generalize Gauss's method to prove the resolvability of other equations, invented what came to be known as the Galois fields and observed that the same methods that were used to prove the existence of primitive roots modulo p could also be used to establish the existence of primitive elements in his fields (Theorem 7.17).

The identification of primitive elements is an issue that puzzled Euler. Both Gauss and Galois later introduced some methodology into this question. Since this would take us outside the scope of this text we merely point out that trial and error can always be used to locate primitive elements in relatively small Galois fields. Thus, the Galois imaginaries α and β of Section 7.1 are clearly primitive elements of $\text{GF}(2, x^2 + x + 1)$ and $\text{GF}(2, x^3 + x^2 + 1)$, respectively, whereas γ is not a primitive element of

$$\text{GF}(2, x^4 + x^3 + x^2 + x + 1)$$

since $\gamma^5 = 1$. However, the element $1 + \gamma$ is a primitive element of this latter field. To see this, note that by Proposition 7.7 and Theorem 7.11, $\phi(1 + \gamma)$ is a divisor of 15. However, by the Binomial Theorem (Theorem 6.2)

$$(1 + \gamma)^3 = 1 + 3\gamma + 3\gamma^2 + \gamma^3 = 1 + \gamma + \gamma^2 + \gamma^3 \neq 1$$

and

$$\begin{aligned}(1 + \gamma)^5 &= 1 + 5\gamma + 10\gamma^2 + 10\gamma^3 + 5\gamma^4 + \gamma^5 = 1 + \gamma + \gamma^4 + \gamma^5 \\ &= 1 + \gamma + (1 + \gamma + \gamma^2 + \gamma^3) + 1 = 1 + \gamma^2 + \gamma^3 \neq 1.\end{aligned}$$

Hence, $\text{o}(1 + \gamma) = 15$ and so $1 + \gamma$ is indeed a primitive element of $\text{GF}(2, x^4 + x^3 + x^2 + x + 1)$. We also note in passing that once it is known that a certain element ζ is a primitive element of some Galois field, then it follows from Proposition 7.7 that the other primitive elements of this field are the powers ζ^m where m is relatively prime to $p^\nu - 1$.

Exercises 7.3

List all the primitive elements of the fields in Exercises 7.3.1 to 7.3.9.

1. $\text{GF}(2, x^2 + x + 1)$
2. $\text{GF}(2, x^3 + x^2 + 1)$
3. \mathbb{Z}_5
4. \mathbb{Z}_{17}
5. $\text{GF}(2, x^4 + x + 1)$
6. $\text{GF}(2, x^4 + x^3 + x^2 + x + 1)$
7. $\text{GF}(3, x^2 + x + 2)$
8. $\text{GF}(5, x^2 + 4x + 2)$
9. $\text{GF}(7, x^2 + 6x + 3)$
10. For any element $a \in \text{GF}(p, P(x))$, let $r(a)$ denote the number of distinct elements b in $\text{GF}(p, P(x))$ such that $b^{p-1} = a$. Prove that if $a \neq 0$, then $r(a) = 0$ or $p - 1$.
11. Let $\text{GF}(p, P(x))$ be any Galois field, and let r be a positive integer such that $r \not\equiv 0 \pmod{p^\nu - 1}$, where ν is the degree of $P(x)$. Prove that the sum of the r -th powers of the elements of $\text{GF}(p, P(x))$ is zero.
12. Prove that the product of all the primitive elements of $\text{GF}(p, P(x))$ is 1, unless $\text{GF}(p, P(x))$ is \mathbb{Z}_3 , in which case this product is 2.
13. Prove that if ζ is a primitive element of $\text{GF}(p, P(x))$ where $P(x)$ has degree ν , then

$$x^{p^\nu} - 1 = \prod_{i=1}^{\nu} (x - \zeta^i).$$

14. Prove that for every prime p , the polynomial $x^p - x - 1$ is irreducible over \mathbb{Z}_p .

7.4 On the Variety of Galois Fields

We now know that for every polynomial $P(x)$ that is irreducible over \mathbb{Z}_p there is a corresponding Galois field $\text{GF}(p, P(x))$. This observation made it possible to construct a variety of new fields each of which contains p^ν elements, where p is some prime and ν is some positive integer. It is appropriate at this point to address the issue of classifying these new structures. There are two questions that every experienced mathematician would ask in this context. Given any such p and ν , does there exist a Galois field of order p^ν ? Given any two such Galois fields, when are they in fact one and the same? Unfortunately, the complete resolutions of these questions lie beyond the bounds of this book and the following discussion provides only informal answers.

One way of proving that a field of order p^ν exists is to display a polynomial of degree ν that is irreducible over \mathbb{Z}_p . This task is not easy. In fact, it turns out to be easier to count the number of all such polynomials than to produce even one. Exercises 6.2.6, 6.2.7, and 7.3.14 deal with some special cases of this issue.

Next, we turn to the second question and reexamine the first Galois field constructed in this chapter, $\text{GF}(2, x^2 + x + 1)$. This field was constructed by stipulating that α is a zero of the polynomial $x^2 + x + 1$ over \mathbb{Z}_2 and then extracting some arithmetical consequences. However, our experience with all the previously constructed fields, namely, the rationals, reals, complex, and modulo p arithmetic, leads us to expect any quadratic to have two zeroes. Should we therefore stipulate the existence of another zero α' of $x^2 + x + 1$ over \mathbb{Z}_2 and then proceed to create its Galois field? This is, of course, unnecessary, since α' gives rise to a field that behaves exactly like that associated with α , except that each occurrence of α in the latter will be replaced by an α' . It turns out that the redundancy goes even deeper. The element α' is already in $\text{GF}(2, x^2 + x + 1)$. In fact $\alpha' = \alpha^2$, for

$$(\alpha^2)^2 + \alpha^2 + 1 = \alpha^4 + \alpha^2 + 1 = \alpha + \alpha^2 + 1 = 0.$$

A similar phenomenon occurs in $\text{GF}(2, x^3 + x^2 + 1)$ whose Galois imaginary was denoted by β . Note that

$$(\beta^2)^3 + (\beta^2)^2 + 1 = \beta^6 + \beta^4 + 1 = (\beta + \beta^2) + (1 + \beta + \beta^2) + 1 = 0 \quad \text{over } \mathbb{Z}_2,$$

implying that β^2 is another zero of $x^3 + x^2 + 1$ in this field. Cubics, however, can have up to three zeroes, so we can expect yet another zero of this polynomial. The element β^3

fails to be such a zero since

$$(\beta^3)^3 + (\beta^3)^2 + 1 = \beta^9 + \beta^6 + 1 = \beta^2 + (\beta + \beta^2) + 1 = 1 + \beta \neq 0 \quad \text{over } \mathbb{Z}_2.$$

However, β^4 turns out to be the third zero of $x^3 + x^2 + 1$ since

$$(\beta^4)^3 + (\beta^4)^2 + 1 = \beta^{12} + \beta^8 + 1 = \beta^5 + \beta + 1 = (1 + \beta) + \beta + 1 = 0 \quad \text{over } \mathbb{Z}_2.$$

With hindsight, we could have argued as follows. Assuming that β was a zero of $x^3 + x^2 + 1$ over \mathbb{Z}_2 it was shown that β^2 was also such a zero. Hence, beginning with the fact that β^2 is a zero of this polynomial, it follows that its square, namely $(\beta^2)^2 = \beta^4$ should also be such a zero, as is indeed the case. Note that the square of β^4 is $\beta^8 = \beta$ which is also a zero of $x^3 + x^2 + 1$, but not a new one.

Let us examine another example before announcing the general principle. Consider the polynomial $x^2 + x + 2$ which is irreducible over \mathbb{Z}_3 , and whose cyclic table was constructed in Section 7.1 in terms of the Galois imaginary σ . Since

$$(\sigma^2)^2 + \sigma^2 + 2 = \sigma^4 + \sigma^2 + 2 = 2 + (1 + 2\sigma) + 2 = 2 + 2\sigma \neq 0,$$

it follows that σ^2 is not another zero of $x^2 + x + 2$ over \mathbb{Z}_3 . However,

$$(\sigma^3)^2 + \sigma^3 + 2 = \sigma^6 + \sigma^3 + 2 = (2 + \sigma) + (2 + 2\sigma) + 2 = 0,$$

and hence σ^3 is the other zero of $x^2 + x + 2$. So, of course, is $(\sigma^3)^3 = \sigma^9 = \sigma$. The pattern is clear and is formulated in Proposition 7.19 below.

Lemma 7.18 If $a, b, c, \dots \in \text{GF}(p, P(x))$, then $(a + b + c + \dots)^p = a^p + b^p + c^p + \dots$.

Proof. Inasmuch as the Binomial Theorem (Theorem 6.2) holds in arbitrary fields, and since, as was argued in the proof of Proposition 5.14, $\binom{p}{k} \equiv 0 \pmod{p}$ for $0 < k < p$, it is seen that $(a + b)^p = a^p + b^p$ for any $a, b \in \text{GF}(p, P(x))$. The lemma now follows by an easy induction argument. ■

Proposition 7.19 If α is a zero of the polynomial $P(x)$ over \mathbb{Z}_p , then so are

$$\alpha, \alpha^p, \alpha^{p^2}, \alpha^{p^3}, \dots$$

zeroes of $P(x)$.

Proof. It clearly suffices to show that if α is a zero of $P(x)$, then so is α^p . Suppose

$$P(x) = a_0 x^n + a_1 x^{n-1} + \cdots + a_{n-1} x + a_n, \quad \text{with } a_0, \dots, a_n \in \mathbb{Z}_p.$$

If α is any zero of $P(x)$, then it follows from the Lemma 7.18 and Fermat's Theorem (Theorem 5.15) that

$$\begin{aligned} P(\alpha^p) &= a_0 (\alpha^p)^n + a_1 (\alpha^p)^{n-1} + \cdots + a_{n-1} \alpha^p + a_n \\ &= a_0^p (\alpha^n)^p + a_1^p (\alpha^{n-1})^p + \cdots + a_{n-1}^p \alpha^p + a_n^p \\ &= (a_0 \alpha^n + a_1 \alpha^{n-1} + \cdots + a_{n-1} \alpha + a_n)^p = 0^p = 0. \end{aligned}$$

Thus, if α is any zero of $P(x)$ so is α^p . ■

When $P(x)$ is not irreducible, the list given in the statement of the above proposition need not contain all of its zeroes (Exercise 7.1.15). When $P(x)$ is irreducible, a stronger claim can be made. We state this without justification, and relegate the proof to Exercises 7.1.7 to 7.1.9.

Proposition 7.20 If α is a zero of the irreducible polynomial $P(x)$ of degree ν over \mathbb{Z}_p , then

$$\alpha, \alpha^p, \alpha^{p^2}, \alpha^{p^3}, \dots, \alpha^{p^{\nu-1}}$$

are all of its zeroes.

We have argued that Galois fields that arise from different imaginaries that correspond to the same irreducible polynomial are in fact one and the same. Surprisingly, a similar phenomenon occurs when the Galois imaginaries correspond to different irreducible polynomials that have the same degree. Such is the case for the fields $\text{GF}(2, x^3 + x^2 + 1)$ and $\text{GF}(2, x^3 + x + 1)$ both of which have $2^3 = 8$ elements. The cyclic table of the first of these appears in Section 7.1, and that of the second is displayed in Table 7.2 (Exercise 7.1.1). These two tables are quite different. Nevertheless, these two fields are really one and the same. To see this observe that β^3 of $\text{GF}(2, x^3 + x^2 + 1)$ is also a zero of $x^3 + x + 1$, the same polynomial that gave rise to the Galois imaginary τ . For,

$$(\beta^3)^3 + \beta^3 + 1 = \beta^9 + \beta^3 + 1 = \beta^2 + \beta^3 + 1 = 0.$$

Thus, having constructed the field $\text{GF}(2, x^3 + x^2 + 1)$, there is no call for associating a new Galois imaginary τ to the polynomial $x^3 + x + 1$ which is irreducible over \mathbb{Z}_2 . The

$\tau^0 = 1$	$\tau^4 = \tau + \tau^2$
$\tau^1 = \tau$	$\tau^5 = 1 + \tau + \tau^2$
$\tau^2 = \tau^2$	$\tau^6 = 1 + \tau^2$
$\tau^3 = 1 + \tau$	$\tau^7 = 1$

Table 7.2 The cyclic table of $\text{GF}(2, x^3 + x^2 + 1)$

element β^3 of $\text{GF}(2, x^3 + x^2 + 1)$ is already a zero of this polynomial, as are $(\beta^3)^2 = \beta^6$ and $(\beta^6)^2 = \beta^5$. In other words, there is an unanticipated redundancy in the process Galois used to create his fields. This is what Galois had in mind when he wrote in the above quoted passage

We are concerned here with the classification of these imaginaries and their minimization.

It is very tempting at this point to argue as follows. If $P(x)$ and $Q(x)$ are irreducible polynomials of degree ν over \mathbb{Z}_p , then, by Corollary 7.15 both $\text{GF}(p, P(x))$ and $\text{GF}(p, Q(x))$ consist of all the zeroes of $x^{p^\nu} - x$, and consequently, these two fields should be one and the same. The flaw in this argument is exposed by the application of an analogous argument to the polynomial $x^2 + 1$. This polynomial has zeroes $\{2, 3\}$ in \mathbb{Z}_5 and $\{4, 13\}$ in \mathbb{Z}_{17} , yet we cannot conclude that $\{2, 3\} = \{4, 13\}$ in any sense. Similarly, the zeroes that $x^{p^\nu} - x$ has in $\text{GF}(p, P(x))$ have, in principle, nothing to do with its zeroes in $\text{GF}(p, Q(x))$. The faultiness of this argument notwithstanding, its conclusion happens to be valid. We state it informally below since the careful formulation and the proof of this theorem lie beyond the scope of this text.

Theorem 7.21 For any prime number p and any positive integer ν there is exactly one Galois field containing p^ν elements.

The field whose existence is guaranteed by this theorem is denoted by $\text{GF}(p^\nu)$. Accordingly,

$$\text{GF}(2, x^3 + x^2 + 1) = \text{GF}(2, x^3 + x + 1) = \text{GF}(2^3)$$

and $\text{GF}(3, x^2 + x + 2) = \text{GF}(3^2)$.

We conclude this section with another aspect of primitivity. A polynomial $P(x)$ of degree ν is said to be *primitive* over \mathbb{Z}_p if it is irreducible over \mathbb{Z}_p and its associated Galois imaginary ξ has order $p^\nu - 1$. In other words, $P(x)$ is primitive over \mathbb{Z}_p when it is irreducible over \mathbb{Z}_p and its Galois imaginary ξ is also primitive. Since the construction of the cyclic table of $\text{GF}(p, P(x))$ depends only on $P(x)$, rather than the specific ξ , it

follows that either all the zeroes of $P(x)$ are primitive or none of them is primitive. Thus the polynomial $x^3 + x^2 + 1$ is primitive over \mathbb{Z}_2 , whereas $x^4 + x^3 + x^2 + x + 1$ is not. Finding a primitive polynomial is not an easy task. However, once such a polynomial has been found, it is easy to find all the other primitive polynomials of the same degree. Consider the polynomial $x^2 + x + 2$ which is known to be primitive over \mathbb{Z}_3 . If σ is its associated Galois imaginary, then, by Proposition 7.7, the elements σ^3 , σ^5 , and σ^7 are all the other primitive elements of $\text{GF}(3, x^2 + x + 2)$. By Proposition 7.19, σ and σ^3 are the zeroes of the same irreducible polynomial, as are σ^5 and $(\sigma^5)^3 = \sigma^{15} = \sigma^7$. Thus the monic primitive quadratic polynomials over \mathbb{Z}_3 are

$$(x - \sigma)(x - \sigma^3) = x^2 - (\sigma + \sigma^3)x + \sigma\sigma^3 = x^2 - (\sigma + 2\sigma + 2)x + \sigma^4 = x^2 + x + 2$$

and

$$(x - \sigma^5)(x - \sigma^7) = x^2 - (\sigma^5 + \sigma^7)x + \sigma^5\sigma^7 = x^2 - (2\sigma + \sigma + 1)x + \sigma^{12} = x^2 + 2x + 2.$$

Exercises 7.4

1. Use the fact that $x^2 + 4x + 2$ is a primitive polynomial to determine all the monic primitive quadratic polynomials over \mathbb{Z}_5 .
2. Verify that the polynomial $x^2 - x + 3$ is primitive over \mathbb{Z}_7 . Write out its cyclic table and use it to find all the other seven monic quadratic polynomials that are primitive over \mathbb{Z}_7 .
3. Determine all the monic primitive cubic polynomials over \mathbb{Z}_3 .
4. Prove that if $2^\nu - 1$ is a prime number, then every polynomial of degree ν that is irreducible over \mathbb{Z}_2 is also primitive.
5. Prove that if p is a prime other than 2 and $P(x)$ is an irreducible polynomial over \mathbb{Z}_p of degree greater than 1, then $\text{GF}(p, (P(x)))$ has some nonprimitive elements besides 0 and 1.
6. Suppose the polynomial $a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n$ is primitive over a field F and $a_0, a_n \neq 0$. Prove that the polynomial $a_nx^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$ is also primitive over F .

7. Prove that for every element α of $\text{GF}(p, P(x))$ there is a positive integer k such that

$$\alpha, \alpha^p, \alpha^{p^2}, \dots, \alpha^{p^{k-1}}$$

are all distinct, and $\alpha^{p^k} = \alpha$.

8. For any $\alpha \in \text{GF}(p, P(x))$, let

$$M_\alpha(x) = (x - \alpha)(x - \alpha^p)(x - \alpha^{p^2}) \cdots (x - \alpha^{p^{k-1}})$$

where k is as defined in Exercise 7.4.7. Prove that $M_\alpha(x) \in \mathbb{Z}_p[x]$.

9. Prove Proposition 7.20.

By Proposition 7.20, any polynomial that has α as its zero is divisible by the polynomial $M_\alpha(x)$ of Exercise 7.4.8. The polynomial $M_\alpha(x)$ is therefore called the (monic) *minimal polynomial* of α . Find the monic minimal polynomials for all the elements of each of the fields in Exercises 7.4.10 to 7.4.14.

10. $\text{GF}(2, x^3 + x + 1)$ 12. $\text{GF}(3, x^2 + 2x + 2)$ 14. $\text{GF}(7, x^2 - x + 3)$

11. $\text{GF}(2, x^4 + x + 1)$ 13. $\text{GF}(5, x^2 + 4x + 2)$

15. Show that the conclusion of Proposition 7.20 need not be valid if $P(x)$ is reducible over \mathbb{Z}_p .

Chapter Summary

Working by analogy with the complex numbers, Galois created a host of new fields, now bearing his name. The nonzero elements of these Galois fields are also roots of unity and as such have orders whose properties are indistinguishable from the orders of the complex and modular roots of unity. We proved the Primitive Element Theorem for these (and the modular) roots of unity. It was also pointed out that these fields are subject to some subtle relationships, since seemingly different fields may turn out, upon careful examination, to be identical.

Chapter Review Exercises

Mark the following true or false.

1. Let $P(x)$ be an irreducible quadratic over \mathbb{Z}_7 , and let α be the associated Galois imaginary. Then there exist $a, b \in \mathbb{Z}_7$ such that $(3 + 4\alpha)(a + b\alpha) = 1$.
2. Let $P(x)$ be an irreducible quadratic over \mathbb{Z}_7 , and let α be the associated Galois imaginary. Then there exist $a, b \in \mathbb{Z}_7$ such that $(2 + 4\alpha)(a + b\alpha) = 1 + 5\alpha$.
3. Let $P(x)$ be an irreducible cubic over \mathbb{Z}_3 . Then $\text{GF}(3, P(x))$ has 30 elements.
4. Let $P(x)$ be an irreducible polynomial over \mathbb{Z}_p . Then the equation $P(x) = 0$ has a solution in $\text{GF}(p, P(x))$.
5. Let $P(x)$ be an irreducible polynomial of degree 4 over \mathbb{Z}_5 . Then every element of $\text{GF}(4, P(x))$ is a zero of the polynomial $x^{600} + x^{25} + x^7 + x$.
6. The polynomial $x^2 + 1$ is primitive over \mathbb{Z}_2 .
7. There is an element r in \mathbb{Z}_{31} such that the sequence $1, r, r^2, \dots \pmod{31}$ contains all the nonzero elements of \mathbb{Z}_{31} .
8. Let $\alpha \in \text{GF}(2, x^5 + x^2 + 1)$. If $P(x) \in \mathbb{Z}_2[x]$ and $P(\alpha) = 0$, then $P(\alpha^2) = 0$.

New Terms

cyclic table, 135

derivative, 154

Galois field, 135

Galois imaginaries, 132

Galois polynomial, 142

minimal polynomial, 152

order, 136

primitive, 135, 150

Supplementary Exercises

1. Write a computer script that lists the primitive monic polynomials of degree d .
2. Prove that for every positive integer d and for every prime p , there is a primitive polynomial of degree d over \mathbb{Z}_p .
3. Find a formula for the number of monic irreducible polynomials of degree d over \mathbb{Z}_p .
4. Write a computer script that will solve any polynomial equation $Q(x) = 0$ over any Galois field $\text{GF}(p, P(x))$.
5. If F is any field and

$$P(x) = a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n \in F[x],$$

then the *derivative* of $P(x)$ is defined as the polynomial

$$P'(x) = na_0x^{n-1} + (n-1)a_1x^{n-2} + \cdots + a_{n-1} \in F[x].$$

Prove that this derivative has the following properties:

- (a) $[P(x) + Q(x)]' = P'(x) + Q'(x)$ for any $P(x), Q(x) \in F[x]$.
 - (b) $[cP(x)]' = cP'(x)$ for any $c \in F$ and $P(x) \in F[x]$.
 - (c) $[P(x)Q(x)]' = P'(x)Q(x) + P(x)Q'(x)$ for any $P(x), Q(x) \in F[x]$.
 - (d) $P(x)$ has repeated zeroes in F only if the greatest common divisor of $P(x)$ and $P'(x)$ has degree at least 1.
6. Let p be any prime and n any positive integer. Prove that the polynomial $x^n - 1$ has repeated zeroes in \mathbb{Z}_p if and only if p is a factor of n .
 7. For how many primes p is 10 a primitive root modulo p ?

Chapter 8



PERMUTATIONS

MOTIVATED BY Lagrange's solution of the quartic equation we consider the general question of what happens to a multivariable function when its variables are permuted. Some surprising results are derived and these lead us to a deeper examination of the notion of a permutation.

8.1 Permuting the Variables of a Function I

The key to Lagrange's solution of the general quartic equation (Section 6.5) was the fact that when the variables of the polynomial $x_1x_2 + x_3x_4$ are permuted in all the possible ways so as to produce the 24 polynomials

$$\begin{array}{llll} x_1x_2 + x_3x_4, & x_1x_2 + x_4x_3, & x_1x_3 + x_2x_4, & x_1x_3 + x_4x_2, \\ x_1x_4 + x_2x_3, & x_1x_4 + x_3x_2, & x_2x_1 + x_3x_4, & x_2x_1 + x_4x_3, \\ x_2x_3 + x_1x_4, & x_2x_3 + x_4x_1, & x_2x_4 + x_1x_3, & x_2x_4 + x_3x_1, \\ x_3x_1 + x_2x_4, & x_3x_1 + x_4x_2, & x_3x_2 + x_1x_4, & x_3x_2 + x_4x_1, \\ x_3x_4 + x_1x_2, & x_3x_4 + x_2x_1, & x_4x_1 + x_2x_3, & x_4x_1 + x_3x_2, \\ x_4x_2 + x_1x_3, & x_4x_2 + x_3x_1, & x_4x_3 + x_1x_2, & x_4x_3 + x_2x_1, \end{array}$$

it follows from the commutativity of addition and multiplication that in fact only three distinct polynomials emerge, namely, $x_1x_2 + x_3x_4$, $x_1x_3 + x_2x_4$, and $x_1x_4 + x_2x_3$. In general, when two polynomials $P(x, y, z, \dots)$ and $Q(x, y, z, \dots)$ can be obtained from each other by permuting their variables, these polynomials are said to be *variants* of each other. Thus, $x_1x_2 + x_3x_4$ has the 24 variants listed above, whereas $x_1x_2 + x_3$ has the six variants $x_1x_2 + x_3$, $x_2x_1 + x_3$, $x_1x_3 + x_2$, $x_3x_1 + x_2$, $x_2x_3 + x_1$, and $x_3x_2 + x_1$. Two

variants are said to be *distinct variants* if they differ as functions over \mathbb{C} . Intuition is a fairly reliable guide in this context. When necessary, however, appropriate substitutions can be used to verify distinctness. Thus, the substitution of 1, 1, 0, 0 for x_1, x_2, x_3, x_4 , respectively, proves that the variants $x_1x_2 + x_3x_4$ and $x_1x_3 + x_2x_4$ are distinct, since they assume the respective values 1 and 0. Lagrange's observation can now be phrased as the polynomial $x_1x_2 + x_3x_4$ has three distinct variants. Similarly, the polynomial $x_1x_2 + x_3$ also has three distinct variants. A function that has no distinct variants is said to be an *invariant function*. The polynomials x_1x_2 and $x_1 + x_2 + x_3$ as well as all the elementary symmetric polynomials of Section 6.4 are examples of invariant functions.

Since Lagrange's solution of the quartic equation hinges on the existence of a polynomial of four variables with three distinct variants, it is reasonable, when attempting the solution of the fifth-degree equation, to look for polynomials in five variables that have four (or perhaps three) distinct variants. Surprisingly, such polynomials do not exist.

Let us digress here and change the point of view somewhat. The five-variable polynomial $x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2$ is clearly invariant. On the other hand, the polynomial $x_1 + x_2 + x_3 + x_4 - x_5$ has five distinct variants, namely, itself and the four polynomials

$$\begin{aligned} x_1 + x_2 + x_3 + x_5 - x_4, & \quad x_1 + x_2 + x_4 + x_5 - x_3, \\ x_1 + x_3 + x_4 + x_5 - x_2, & \quad x_2 + x_3 + x_4 + x_5 - x_1. \end{aligned}$$

Similarly, the polynomial $x_1x_2 + x_3x_4x_5$ has $\binom{5}{2} = 10$ distinct variants, namely, itself and the nine polynomials

$$\begin{aligned} x_1x_3 + x_2x_4x_5, & \quad x_1x_4 + x_2x_3x_5, & \quad x_1x_5 + x_2x_3x_4, \\ x_2x_3 + x_1x_4x_5, & \quad x_2x_4 + x_1x_3x_5, & \quad x_2x_5 + x_1x_3x_4, \\ x_3x_4 + x_1x_2x_5, & \quad x_3x_5 + x_1x_2x_4, & \quad x_4x_5 + x_1x_2x_3. \end{aligned}$$

However, is there a function of these five variables that has four distinct variants? A formal proof of the nonexistence of such a function is offered below in Corollary 8.11. The need for such a proof is underscored by the fact that the beginner's search for functions of five variables that has two distinct variants is very likely also to meet with failure. Such functions, however, do exist, and a method for constructing them is suggested in Exercise 8.1.36. A complete proof of the existence of such two-variant functions is offered in Proposition 8.14.

That there are no functions of five variables that have four (or three) distinct variants was first recognized by Paolo Ruffini, who incorporated this observation into his unsuccessful attempt to prove the unsolvability of the general quintic equation by radicals. Ruffini's theorem regarding the number of distinct variants that a function of five variables can have was generalized in 1815 by Cauchy to the statement that appears as Theorem 8.10 below. This proposition was incorporated by Abel into his groundbreaking proof of the unsolvability of the general quintic equation by radicals. In 1847 Cauchy returned to this topic and proved that if a function on $n \geq 5$ variables has fewer than n distinct variants, then that function has only one or two distinct variants. Since this stronger version turned out to play no special role in the evolution of the theory of algebraic resolvability of equations, it is mentioned without proof. We shall, however, prove Cauchy's 1815 theorem after providing some basic theoretical information about permutations.

Exercises 8.1

Find the number of distinct variants that the functions in Exercises 8.1.1 to 8.1.35 have.

- | | | |
|--|--|---------------------------|
| 1. $x_1 + x_2$ | 5. $x_1/x_2 + x_2/x_1$ | 9. $x_1 x_2^2$ |
| 2. $x_1 - x_2$ | 6. $\sin(x - y)$ | 10. $x_1^3 x_2^3 x_3^3$ |
| 3. $(x_1 - x_2)^2$ | 7. $\cos(x - y)$ | 11. $x_1 x_2 x_3^2$ |
| 4. x_1/x_2 | 8. $x_1 + 5x_2$ | 12. $(x_1 + x_2 - x_3)^2$ |
| 13. $(x_1 + x_2)(x_1 + x_3)(x_2 + x_3)$ | 20. $x_1 x_2 x_3^5 x_4^5$ | |
| 14. $(x_1 - x_2)(x_1 - x_3)(x_2 - x_3)$ | 21. $x_1 x_2 - x_3 x_4$ | |
| 15. $x_1 x_2 + x_3$ | 22. $(x_1 x_2 - x_3 x_4)^2$ | |
| 16. $x_1/x_2 + x_3$ | 23. $(x_1 + x_2)/(x_3 x_4)$ | |
| 17. $x_1 x_2^2 x_3^2$ | 24. $(x_1 - x_2)^2 + (x_3 - x_4)^2$ | |
| 18. $x_1 x_2 x_3 x_4$ | 25. $x_1 x_2^2 x_3^3 x_4^4$ | |
| 19. $x_1 x_2 x_3 x_4^3$ | 26. $(x_1 - x_2)(x_1 - x_3)(x_2 - x_3)x_4$ | |
| 27. $(x_1 + x_2)(x_1 + x_3)(x_1 + x_4)(x_2 + x_3)(x_2 + x_4)(x_3 + x_4)$ | | |
| 28. $(x_1 - x_2)(x_1 - x_3)(x_1 - x_4)(x_2 - x_3)(x_2 - x_4)(x_3 - x_4)$ | | |

29. $x_1 x_2 x_3 x_4 x_5$ 30. $x_1 x_2 x_3 x_4 / x_5$ 31. $(x_1 x_2 x_3 + x_4) / x_5$
 32. $(x_1 x_2 x_3) / (x_4 x_5)$ 34. $(x_1 - x_2)(x_1 - x_3)(x_2 - x_3)x_4 x_5$
 33. $(x_1 x_2 + x_3 x_4)x_5$ 35. $x_1 x_2^2 x_3^3 x_4^4 x_5^5$

36. Use the answers of Exercise 8.1.14 and Exercise 8.1.28 to create a function of five variables that has two distinct variants.

Prove that for any positive integer n there exists a function of n variables that has the number of distinct variants that is specified in Exercises 8.1.37 to 8.1.43.

37. 1 40. $3\binom{n}{3}$, $n \geq 3$ 43. $k\binom{n}{k}$, $n \geq k \geq 1$
 38. n 41. $3\binom{n}{4}$, $n \geq 4$
 39. $\binom{n}{2}$, $n \geq 2$ 42. $\binom{n}{k}$, $n \geq k \geq 0$

44. Show that for any positive integer n there is a function of n variables such that every two of its variants are distinct.

8.2 Permutations

In the previous section we had several occasions to shuffle some variables and to observe the effect that this transformation had on a function of these variables. We now focus on the shuffles themselves. The mathematical name for such a shuffle is a *permutation*. More formally, a permutation of a set S is a function σ that assigns to each element x of S an element $y = \sigma(x)$ of S so that

- (a) if x_1 and x_2 are distinct elements of S , then $\sigma(x_1) \neq \sigma(x_2)$; and
 (b) if y is any element of S , then there is an element x in S such that $y = \sigma(x)$.

We note in passing that when the underlying set S is finite these two conditions are equivalent (Exercise 8.2.35) and so only one need be verified. The *identity permutation* that transforms each element to itself is denoted by Id . In the earlier days of permutation theory, each permutation was written as an array of two rows. The first of these rows listed the elements of S in some natural order, and the second row listed the corresponding values of σ . Thus, the array

$$\begin{pmatrix} x_1 & x_2 & x_3 & x_4 \\ x_2 & x_3 & x_4 & x_1 \end{pmatrix}$$

was used to denote the permutation σ such that

$$\sigma(x_i) = x_{i+1} \quad (\text{addition modulo 4})$$

and whose effect is to convert the polynomial $x_1x_2 + x_3x_4$ to the polynomial $x_2x_3 + x_4x_1$. Similarly, the permutation that interchanges x_1 with x_3 and also interchanges x_2 with x_4 was denoted by

$$\begin{pmatrix} x_1 & x_2 & x_3 & x_4 \\ x_3 & x_4 & x_1 & x_2 \end{pmatrix}.$$

The letter x above serves merely as a place holder and it is more efficient to eliminate it. Thus, we shall generally restrict our attention to permutations on a set $S = 1, 2, \dots, n$, and the above two permutations can be denoted by

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 1 & 2 \end{pmatrix},$$

respectively. This notation will be further improved, but first we pause to count the permutations of a given set.

Proposition 8.1 For every positive integer n the number of permutations of the set $S = 1, 2, 3, \dots, n$ is $n! = 1 \cdot 2 \cdot 3 \cdots n$.

Proof. Let

$$\begin{pmatrix} 1 & 2 & 3 & \dots & k & \dots & n \\ a & b & c & \dots & h & \dots & j \end{pmatrix}$$

denote an arbitrary permutation of $S = \{1, 2, 3, \dots, n\}$. Then the symbol a can be replaced by any of the n elements of S . Once a has been chosen, b can be replaced by any of the $n-1$ elements of $S - \{a\}$. Once b has been chosen, c can be replaced by any of the $n-2$ elements of $S - \{a, b\}$. Proceeding in the same manner, it is clear that the second row of this arbitrary permutation can be filled out in $n(n-1)(n-2)\cdots 1 = n!$ ways so as to define a bona fide permutation of S . ■

Like all functions, permutations can be composed, and this composition is associative (Exercise 8.2.29). If ρ and σ are two permutations of the set S , then their composition is denoted by either $\rho \circ \sigma$ or simply by their juxtaposition $\rho\sigma$ where $\rho\sigma(a) = \rho(\sigma(a))$. Thus, if

$$\rho = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 1 & 4 & 2 \end{pmatrix} \quad \text{and} \quad \sigma = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \end{pmatrix},$$

then

$$\rho \circ \sigma(1) = \rho\sigma(1) = \rho(\sigma(1)) = \rho(2) = 1,$$

$$\rho \circ \sigma(2) = \rho\sigma(2) = \rho(\sigma(2)) = \rho(3) = 4,$$

$$\rho \circ \sigma(3) = \rho\sigma(3) = \rho(\sigma(3)) = \rho(4) = 2,$$

$$\rho \circ \sigma(4) = \rho\sigma(4) = \rho(\sigma(4)) = \rho(1) = 3.$$

In fact,

$$\rho \circ \sigma = \rho\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 2 & 3 \end{pmatrix}.$$

This situation is quite typical and merits an explicit statement and perhaps also a formal proof.

Proposition 8.2 If ρ and σ are permutations of the set S , then so is their composition $\rho\sigma$ a permutation of the set S .

Proof. Suppose x_1 and x_2 are distinct elements of S . Then, since both ρ and σ are known to be permutations, it follows from the definition first that $\sigma(x_1) \neq \sigma(x_2)$ and next that $\rho(\sigma(x_1)) \neq \rho(\sigma(x_2))$, or $\rho\sigma(x_1) \neq \rho\sigma(x_2)$. Thus, the composition $\rho\sigma$ also satisfies the first required property.

Similarly, if y is any element of S , then by definition there exists an element z of S such that $\rho(z) = y$, and, by the same argument, there exists an element x of S such that $\sigma(x) = z$. Combining these two we see that

$$\rho\sigma(x) = \rho(\sigma(x)) = \rho(z) = y,$$

so that the composition $\rho\sigma$ also the second required property. Thus $\rho\sigma$ is a permutation of S . ■

If σ is any permutation, then we define $\sigma^0 = \text{Id}$, $\sigma^1 = \sigma$, $\sigma^2 = \sigma\sigma$, and, if k is any positive integer, σ^k is the composition of k σ 's. Accordingly, if

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 3 & 4 & 5 & 1 \end{pmatrix}, \quad \text{then} \quad \sigma^2 = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 1 & 2 \end{pmatrix} \quad \text{and} \quad \sigma^3 = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 4 & 5 & 1 & 2 & 3 \end{pmatrix}.$$

We now describe yet another, more efficient, way of writing down permutations of finite sets. We first define a *cycle*, or a *cyclic permutation*, as a permutation of the form

$$\begin{pmatrix} a & b & c & \dots & g & h \\ b & c & d & \dots & h & a \end{pmatrix}$$

and agree to write it in any of the forms

$$(a \ b \ c \ \dots \ g \ h) = (b \ c \ \dots \ g \ h \ a) = (c \ d \ \dots \ h \ a \ b) = \dots = (h \ a \ b \ \dots \ g).$$

If k is any positive integer, then a k -*cycle* is a cycle that contains k elements. Thus, $(3 \ 5 \ 7)$ is a 3-cycle and $(1 \ 8 \ 7 \ 2 \ 5)$ is a 5-cycle. Suppose next that σ is an arbitrary permutation and a is an arbitrary element of the underlying set S . Consider the sequence

$$\sigma^0(a) = a, \sigma^1(a) = \sigma(a), \sigma^2(a), \sigma^3(a), \dots$$

Since S is finite, this infinite sequence must contain repetitions. Let k be the first exponent for which there exists another exponent $m > k$ such that $\sigma^k(a) = \sigma^m(a)$. The exponent k must in fact be 0, since otherwise we would have

$$\sigma(\sigma^{k-1}(a)) = \sigma^k(a) = \sigma^m(a) = \sigma(\sigma^{m-1}(a))$$

and by the first property of permutations we would have $\sigma^{k-1}(a) = \sigma^{m-1}(a)$, contradicting the minimality of k . Hence the elements of

$$a, \sigma(a), \sigma^2(a), \sigma^3(a), \dots, \sigma^{k-1}(a) \tag{8.3}$$

are all distinct and $\sigma(\sigma^{k-1}(a)) = a$. If we define σ_a to be the cycle

$$(a \ \sigma(a) \ \sigma^2(a) \ \sigma^3(a) \ \dots \ \sigma^{k-1}(a)),$$

then it is clear that σ and σ_a agree on all the elements of List 8.3. If List 8.3 does not exhaust all the elements permuted by σ , let b be an element that does not appear in List 8.3, and let h be the least positive integer such that $\sigma^h(b) = b$. We then define

$$\sigma_b = (b \ \sigma(b) \ \sigma^2(b) \ \sigma^3(b) \ \dots \ \sigma^{h-1}(b)).$$

If this process is repeated until all the elements permuted by σ are exhausted, we have cyclic permutations $\sigma_a, \sigma_b, \dots$ such that

$$\sigma = \sigma_a \sigma_b \cdots \quad (8.4)$$

Thus, for example,

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 9 & 5 & 1 & 3 & 7 & 6 & 2 & 4 & 8 \end{pmatrix} = (1\ 9\ 8\ 4\ 3)(6)(2\ 5\ 7).$$

We shall refer to Equation 8.4 as the *disjoint cycle decomposition* of σ . Note that the order in which the individual cycles in the disjoint cycle form of a permutation are written is arbitrary. Similarly, it is only the cyclic order of the elements that appear in a cycle that is significant. Thus,

$$\begin{aligned} (1\ 2\ 3\ 4)(5\ 6\ 7)(8\ 9) &= (5\ 6\ 7)(1\ 2\ 3\ 4)(8\ 9) \\ &= (8\ 9)(5\ 6\ 7)(1\ 2\ 3\ 4) = (3\ 4\ 1\ 2)(8\ 9)(6\ 7\ 5). \end{aligned}$$

If this process gives rise to a cycle of length 1, that cycle is generally omitted. Accordingly,

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 2 & 4 & 6 & 5 & 1 \end{pmatrix} = (1\ 3\ 4\ 6).$$

It is clear that if $\rho = (a_1\ a_2\ a_3\ \dots\ a_{n-1}\ a_n)$ and $\sigma = (a_n\ a_{n-1}\ \dots\ a_3\ a_2\ a_1)$, then $\rho\sigma = \sigma\rho = \text{Id}$. We say that ρ and σ are *inverses* of each other and write $\sigma = \rho^{-1}$ and $\rho = \sigma^{-1}$. Even when ρ is not necessarily cyclic, it has an inverse and this inverse is easily described. To see this, let $\sigma = \sigma^1\sigma^2\cdots\sigma^{n-1}\sigma^n$ be the disjoint cycle decomposition of σ . If we now set

$$\rho = \sigma_n^{-1}\sigma_{n-1}^{-1}\cdots\sigma_2^{-1}\sigma_1^{-1},$$

then

$$\sigma\rho = \sigma_1\sigma_2\cdots\sigma_{n-1}\sigma_n\sigma_n^{-1}\sigma_{n-1}^{-1}\cdots\sigma_2^{-1}\sigma_1^{-1} = \sigma_1\sigma_2\cdots\sigma_{n-1}\sigma_{n-1}^{-1}\cdots\sigma_2^{-1}\sigma_1^{-1} = \cdots = \text{Id}.$$

Similarly, $\rho\sigma = \text{Id}$. Thus, the inverse of the permutation $(1\ 9\ 8\ 4\ 3)(6)(2\ 5\ 7)$ is the permutation $(7\ 5\ 2)(6)(3\ 4\ 8\ 9\ 1)$.

It is clear from this description of inverses that every permutation has a unique inverse. If we set $\sigma^{-m} = (\sigma^{-1})^m$ for every nonnegative integer, then the powers of permutations obey the usual exponential rules (Exercise 8.2.28). The following lemma, however, is not quite so obvious.

Lemma 8.5 If ρ and σ are permutations of the same set, then $(\rho\sigma)^{-1} = \sigma^{-1}\rho^{-1}$.

Proof. The proof is a straightforward application of the associativity of the composition of permutations. Note that

$$(\rho\sigma)(\sigma^{-1}\rho^{-1}) = \rho(\sigma\sigma^{-1})\rho^{-1} = \rho \text{Id} \rho^{-1} = \rho\rho^{-1} = \text{Id}$$

and

$$(\sigma^{-1}\rho^{-1})(\rho\sigma) = \sigma^{-1}(\rho^{-1}\rho)\sigma = \sigma^{-1} \text{Id} \sigma = \sigma^{-1}\sigma = \text{Id}.$$

Thus, $\sigma^{-1}\rho^{-1}$ fulfills the requisite conditions for being the inverse of $\rho\sigma$. ■

The analysis of the effects of permutations is often facilitated by factoring them into the composition of “smaller” permutations. A *transposition* is a permutation that interchanges only two elements, leaving all the others fixed. Thus every transposition has the form $(a\ b)$. It is clear that in an informal sense the transpositions are the smallest nontrivial permutations. The equations $(1\ 2\ 3) = (1\ 2)(2\ 3)$ and $(1\ 2\ 3\ 4) = (1\ 2)(2\ 3)(3\ 4)$ are instances of nontranspositions expressed as the composition of transpositions. This is always possible.

Proposition 8.6 Every permutation of a finite set is the composition of some transpositions.

Proof. We already know that every permutation of a finite set is the composition of cyclic permutations, and hence it suffices to show that every cyclic permutation can be expressed as the composition of some transpositions. This, however, is easily accomplished as follows: $(a_1\ a_2\ a_3\ \dots\ a_n) = (a_1\ a_2)(a_2\ a_3)\cdots(a_{n-1}\ a_n)$. ■

Thus,

$$(1\ 5\ 3)(2\ 4\ 8\ 9)(a\ b\ c\ d\ e) = (1\ 5)(5\ 3)(2\ 4)(4\ 8)(8\ 9)(a\ b)(b\ c)(c\ d)(d\ e).$$

It should be stressed that such expressions are not unique as illustrated by the equations

$$\begin{aligned}(1\ 2\ 3) &= (1\ 2)(2\ 3) = (1\ 3)(1\ 2) = (2\ 3)(1\ 3) \\ &= (1\ 2)(3\ 4)(2\ 4)(3\ 4) = (2\ 3)(1\ 3)(2\ 3)(3\ 4)(2\ 4)(3\ 4).\end{aligned}$$

It is clear that if $\rho = (a_1\ a_2\ \dots\ a_k)$ is any cyclic permutation then $\rho^k = \text{Id}$. Consequently, if m is any common multiple of the lengths of $\sigma_1, \sigma_2, \dots, \sigma_k$ in the disjoint cycle factorization of the arbitrary permutation $\sigma = \sigma_1\sigma_2\cdots\sigma_k$, then $\sigma^m = \text{Id}$. Thus,

$$((1\ 2\ 3)(4\ 5))^6 = \text{Id}.$$

The order $o(\sigma)$ of the permutation σ is the least positive integer m such that $\sigma^m = \text{Id}$. The above considerations make it clear that every permutation of a finite set has a finite order. Exercise 8.2.31 asserts that the order of any permutation equals the least common multiple of the lengths of its disjoint cyclic factors.

Exercises 8.2

Express the permutations in Exercises 8.1.1 to 8.1.4 in the disjoint cycle form.

- | | |
|---|---|
| 1. $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 9 & 3 & 4 & 7 & 1 & 5 & 2 & 6 & 8 \end{pmatrix}$ | 3. $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 9 & 8 & 7 & 6 & 5 & 4 & 3 & 2 & 1 \end{pmatrix}$ |
| 2. $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 5 & 3 & 4 & 7 & 1 & 6 & 2 & 9 & 8 \end{pmatrix}$ | 4. $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 9 & 7 & 5 & 3 & 1 & 8 & 6 & 4 & 2 \end{pmatrix}$ |

5. List the permutations of $\{1, 2\}$ in disjoint cycle form.
6. List the permutations of $\{1, 2, 3\}$ in disjoint cycle form.
7. List the permutations of $\{1, 2, 3, 4\}$ in disjoint cycle form.

Given the permutations $\rho = (1\ 2\ 3\ 4)(5\ 6\ 7)(8\ 9)$ and $\sigma = (1\ 9\ 8\ 6\ 5)(2\ 3\ 4\ 7)$, express the permutations in Exercises 8.2.8 to 8.2.14 in disjoint cycle form and compute their orders.

- | | | | |
|-----------------|----------------------------|-----------------------------|----------------------|
| 8. $\rho\sigma$ | 10. $\rho\sigma\rho$ | 12. $\rho\sigma\rho^{-1}$ | 14. $\rho^2\sigma^3$ |
| 9. $\sigma\rho$ | 11. $\rho\sigma\rho\sigma$ | 13. $\sigma\rho\sigma^{-1}$ | |

Express the permutations in Exercises 8.2.15 to 8.2.18 as a composition of transpositions.

$$15. (1\ 2\ 3)(4\ 5\ 6\ 7)(8\ 9)$$

$$17. \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 9 & 3 & 4 & 7 & 1 & 5 & 2 & 6 & 8 \end{pmatrix}$$

$$16. (1\ 4\ 2\ 9)(5\ 6\ 3\ 7)$$

$$18. \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 9 & 1 & 5 & 3 & 7 & 8 & 6 & 2 & 4 \end{pmatrix}$$

19. Prove that if $n \geq 2$, then every permutation of $\{1, 2, \dots, n\}$ can be expressed as the composition of transpositions of the form $(1\ a)$, where $a = 2, 3, \dots, n$.
20. Show that if $n \geq 4$, then every permutation of $\{1, 2, \dots, n\}$ is expressible as the composition of 4-cycles.
21. Show that if $n \geq k \geq 2$ and k is even, then every permutation of $\{1, 2, \dots, n\}$ is expressible as the composition of k -cycles.
22. Prove that if σ and ρ are permutations of $\{1, 2, \dots, n\}$, then $o(\sigma) = o(\rho\sigma\rho^{-1})$.
23. Prove the following:
 - (a) If any cycle of the permutation σ has the form $(a_1\ a_2\ a_3 \dots a_k)$, and if ρ is any other permutation, then $\rho\sigma\rho^{-1}$ has a corresponding cycle of the form

$$(\rho(a_1)\ \rho(a_2)\ \rho(a_3) \dots \rho(a_k)).$$

- (b) For each positive integer k , σ and $\rho\sigma\rho^{-1}$ have the same number of k -cycles.
- (c) If the two permutations σ and τ possess the same number of k -cycles for each positive integer k , then there is a permutation ρ such that $\tau = \rho\sigma\rho^{-1}$.

The number of cycles in the disjoint cycle form of the permutation σ on $\{1, 2, \dots, n\}$ is denoted by $\|\sigma\|$. Thus $\|(1\ 2\ 3)(4\ 5)(6\ 7\ 8\ 9)\| = 3$ if $n = 9$, and $\|\text{Id}\| = n$ regardless of the value of n .

24. Let σ be any permutation and let τ be any transposition on the same set. Prove that $\|\sigma\tau\| = \|\sigma\| \pm 1$.
25. Prove that if σ and ρ are any two permutations, then $\|\rho\sigma\rho^{-1}\| = \|\sigma\|$.
26. Prove that if σ and ρ are any two permutations, then $\|\sigma\rho\| = \|\rho\sigma\|$.
27. Prove that every permutation of $\{1, 2, \dots, n\}$ is expressible as the composition of two permutations of $\{1, 2, \dots, n\}$ that have order at most 2.
28. Prove that if σ is any permutation and m and n are any integers, then $\sigma^m\sigma^n = \sigma^{m+n}$ and $(\sigma^m)^n = \sigma^{mn}$.

29. Let S , T , U , and V be sets, and let f be a function from S to T , g a function from T to U , and h a function from U to V . If the composition $g \circ f$ is defined via $g \circ f(x) = g(f(x))$, prove that $h \circ (g \circ f) = (h \circ g) \circ f$.
30. Find two permutations ρ and σ that have relatively prime orders and for which $o(\rho\sigma) \neq o(\rho)o(\sigma)$.
31. Suppose $\sigma = \sigma_1\sigma_2 \cdots \sigma_k$ is the disjoint cycle form of σ , and suppose that each factor σ_i contains m_i elements of S for $i = 1, 2, \dots, k$. Prove that $o(\sigma)$ is the least common multiple of m_1, m_2, \dots, m_k .
32. Prove that the order of every permutation on $\{1, 2, \dots, n\}$ is a proper factor of $n!$ if $n > 2$.
33. For any integers $k \geq 0$ and $n > 0$, let $s(n, k)$ denote the number of permutations of $\{1, 2, \dots, n\}$ whose disjoint cycle decomposition has exactly k cycles. Prove that if $k \geq 1$ and $n > 1$, then $s(n, k) = s(n-1, k-1) + (n-1)s(n-1, k)$.
34. Show that the average number of cycles in the disjoint cycle decomposition of all the permutations on $\{1, 2, \dots, n\}$ is $1 + 1/2 + 1/3 + \cdots + 1/n$.
35. Prove that if the set S is finite and σ is a function of S into itself, then the following conditions are equivalent:
- If x_1 and x_2 are distinct elements of S , then $\sigma(x_1) \neq \sigma(x_2)$.
 - If y is any element of S , then there is an element x in S such that $y = \sigma(x)$.
36. Show, by means of examples, that when S is an infinite set, neither of the conditions of Exercise 8.2.35 need entail the other.

8.3 Permuting the Variables of a Function II

We now return to the issue of the number of distinct variants of a given function of several variables. We begin by formalizing the notion of interchanging the variables of a function. If $f = f(x_1, x_2, \dots, x_n)$ is any function of n variables, and if σ is any permutation of the indices, then we define

$$\sigma f = f(x_{\sigma(1)}, x_{\sigma(2)}, \dots, x_{\sigma(n)}).$$

If $f = x_1x_2 + x_3x_4$ and σ is the cyclic permutation $(1\ 2\ 3\ 4)$, then $\sigma f = x_2x_3 + x_4x_1$.

Lemma 8.7 If f and g are two functions of the variables x_1, x_2, \dots, x_n such that $f = g$, and if σ is any permutation of $\{1, 2, \dots, n\}$, then $\sigma f = \sigma g$.

The next two observations are easy, but fundamental. Their justification relies on the fact that in the composition $\rho\sigma$ it is σ that acts first on the permuted elements, and its action is followed by that of ρ .

Lemma 8.8 If f is any function of the variables x_1, x_2, \dots, x_n , and if ρ and σ are any two permutations of $\{1, 2, \dots, n\}$, then $(\sigma\rho)f = \sigma(\rho f)$.

Corollary 8.9 If $f = \sigma f$, then $f = \sigma^{-1}f$.

Proof. If $f = \sigma f$, then, by Lemmas 8.7 and 8.8,

$$\sigma^{-1}f = \sigma^{-1}(\sigma f) = (\sigma^{-1}\sigma)f = (\text{Id})f = f. \quad \blacksquare$$

If it so happens that $\sigma f = f$, we say that σ *leaves f unchanged*. Thus, $(1\ 3)(2\ 4)$ leaves $x_1x_2 + x_3x_4$ unchanged. We are now ready to state and prove this section's main theorem.

Theorem 8.10 (Cauchy) Let f be a function of n variables, and let p be any prime such that $p \leq n$. If f has $k < p$ distinct variants, then k is either 1 or 2.

Proof. Let f and $k < p \leq n$ be as in the statement of the theorem. We first show that if σ is any permutation of order p of $\{1, 2, \dots, n\}$, then $\sigma f = f$. Consider the p functions

$$f, \sigma f, \sigma^2 f, \dots, \sigma^{p-1} f.$$

Since f has only $k < p$ different variants, it follows that there exist two distinct integers r and s with $0 \leq r < s < p$ such that

$$\sigma^s f = \sigma^r f \quad \text{or} \quad \sigma^{s-r} f = f.$$

It is therefore clear that $\sigma^{a(s-r)} f = f$ for every integer a . Since $0 < s - r < p$, $s - r$ is relatively prime to p so that there exist integers A and B such that $A(s - r) + Bp = 1$. But then, bearing in mind that $\sigma^p = \text{Id}$,

$$f = \sigma^{A(s-r)} f = \sigma^{-Bp+1} f = \sigma(\sigma^{-Bp} f) = \sigma(\text{Id}^{-B} f) = \sigma f.$$

Thus we have proved that if σ is any permutation of order p , then $\sigma f = f$.

Next we show that the application of any two transpositions to the variables of f also leaves f unchanged. Let α and β be the two specific permutations

$$\alpha = (1\ 2\ 3\ 4 \dots p) \quad \text{and} \quad \beta = (p\ p-1 \dots 4\ 2\ 3\ 1).$$

Since they both have order p , both leave f unchanged. Consequently, their composition

$$\beta\alpha = (1\ 3\ 2) = (3\ 2)(2\ 1)$$

also leaves f unchanged. By a similar argument, the permutation

$$(2\ 4\ 3) = (4\ 3)(3\ 2)$$

also leaves f unchanged, as must their composition

$$(4\ 3)(3\ 2)(3\ 2)(2\ 1) = (4\ 3)(2\ 1).$$

Since there was nothing special about the choice of the variables x_1, x_2, x_3 , and x_4 , we now know that f is unchanged whenever its variables are permuted by two, or any even number of consecutive transpositions.

Finally, we demonstrate that if ρ and σ are any permutations expressible as the composition of an odd number of transpositions then $\rho f = \sigma f$. Since it is already known that every permutation is expressible as the composition of some number of transpositions, this will conclude the proof of the theorem. Note that it suffices to show that for any such σ , $(1\ 2)f = \sigma f$. However, since σ is the product of an odd number of transpositions, it follows that $(1\ 2)\sigma$ is the product of an even number of transpositions, so that by the previous argument

$$f = (1\ 2)\sigma f$$

and so

$$(1\ 2)f = (1\ 2)(1\ 2)\sigma f = \sigma f.$$

Thus, in general, every variant σf of f equals either f or $(1\ 2)f$, depending on whether σ is expressible as the composition of an even or an odd number of transpositions. Of course, it could happen that f and $(1\ 2)f$ might be equal, in which case f is invariant. In other words, k is either 2 or 1. ■

While the above theorem was first stated and proved by Cauchy in 1815, the proof we gave is based on that which appears in Abel's 1826 memoir. A translation of Abel's proof is presented in Appendix C.

Corollary 8.11 (Ruffini) There exists no function of five variables that has either three or four distinct variants.

Exercises 8.3

For which of the values of n and k in Exercises 8.3.1 to 8.3.26 does there exist a function of n variables that has k distinct variants? Justify your answers.

- | | | |
|-------------------|---------------------|---|
| 1. $n = 1, k = 1$ | 10. $n = 4, k = 4$ | 19. $n = 6, k = 4$ |
| 2. $n = 2, k = 1$ | 11. $n = 5, k = 1$ | 20. $n = 6, k = 6$ |
| 3. $n = 2, k = 2$ | 12. $n = 5, k = 2$ | 21. $n = 7, k = 6$ |
| 4. $n = 3, k = 1$ | 13. $n = 5, k = 3$ | 22. n arbitrary, $k = n$ |
| 5. $n = 3, k = 2$ | 14. $n = 5, k = 4$ | 23. $n \geq 2, k = \binom{n}{2}$ |
| 6. $n = 3, k = 3$ | 15. $n = 5, k = 5$ | 24. $n \geq 3, k = 3\binom{n}{3}$ |
| 7. $n = 4, k = 1$ | 16. $n = 5, k = 10$ | 25. $n \geq 4, k = 3\binom{n}{4}$ |
| 8. $n = 4, k = 2$ | 17. $n = 6, k = 1$ | 26. $n \geq r \geq 0, k = \binom{n}{r}$ |
| 9. $n = 4, k = 3$ | 18. $n = 6, k = 3$ | |

8.4 The Parity of a Permutation

It was seen above that every permutation can be expressed as the composition of transpositions. Moreover, in the last paragraphs of the proof of (Cauchy's) Theorem 8.10, the issue of the parity of the number of factors in this expression became significant. It was noted earlier that any permutation can be factored into transpositions in many ways, as illustrated by the equations

$$\begin{aligned}
 (1\ 2\ 3) &= (1\ 3)(1\ 2) \\
 &= (1\ 3)(1\ 4)(2\ 4)(1\ 4) \\
 &= (1\ 3)(1\ 2)(2\ 4)(1\ 2)(2\ 4)(1\ 4).
 \end{aligned}$$

However, as indicated by this example, the parity of the number of transpositions in any such factorization of a given permutation is fixed, a fact that we now set out to prove.

For every integer $n \geq 2$ we define the *discriminant* Δ_n as the polynomial

$$\Delta_n = (x_1 - x_2)(x_1 - x_3) \cdots (x_1 - x_n)(x_2 - x_3) \cdots (x_{n-1} - x_n).$$

Accordingly,

$$\Delta_2 = x_1 - x_2,$$

$$\Delta_3 = (x_1 - x_2)(x_1 - x_3)(x_2 - x_3),$$

$$\Delta_4 = (x_1 - x_2)(x_1 - x_3)(x_1 - x_4)(x_2 - x_3)(x_2 - x_4)(x_3 - x_4).$$

It will now be demonstrated, as promised in Section 8.1, that for each integer $n \geq 2$ the discriminant Δ_n has two distinct variants. This fact, in turn, will be used to draw some interesting conclusions regarding the parities of permutations. The brunt of the work is contained in the proof of the following observation.

Lemma 8.12 If τ is any transposition, then $\tau \Delta_n = -\Delta_n$.

Proof. Suppose first that $\tau = (i \ i+1)$ with $1 \leq i < n$. The effect of τ on Δ_n is to replace the segment

$$(x_i - x_{i+1})(x_i - x_{i+2}) \cdots (x_i - x_n)(x_{i+1} - x_{i+2}) \cdots (x_{i+1} - x_n)$$

with

$$(x_{i+1} - x_i)(x_{i+1} - x_{i+2}) \cdots (x_{i+1} - x_n)(x_i - x_{i+2}) \cdots (x_i - x_n).$$

Since $(x_{i+1} - x_i) = -(x_i - x_{i+1})$, it follows that for $\tau = (i \ i+1)$ we do indeed have $\tau \Delta_n = -\Delta_n$.

Next let τ be an arbitrary transposition $(a \ b)$ with $a < b$. Since

$$\begin{aligned} (a \ b) &= (a \ a+1 \ a+2 \ \dots \ b-1 \ b)(b-1 \ b-2 \ \dots \ a+1 \ a) \\ &= (a \ a+1)(a+1 \ a+2) \cdots (b-2 \ b-1)(b-1 \ b)(b-1 \ b-2)(b-2 \ b-3) \\ &\quad \cdots (a+2 \ a+1)(a+1 \ a) \quad (8.13) \end{aligned}$$

and since the right-hand side of Equation 8.13 consists of an odd number (specifically, $2(b-a)-1$) transpositions of the form $(i \ i+1)$, it follows from the first part of the proof that in this case too $\tau \Delta_n = -\Delta_n$. ■

A permutation is said to be an *even permutation* or an *odd permutation* if it is expressible as the composition of an even or odd number of transpositions, respectively. Thus, every transposition is necessarily odd whereas every 3-cycle of the form $(a \ b \ c)$ is even since

$(a\ b\ c) = (a\ b)(b\ c)$. Similarly,

$$(1\ 2\ 3)(4\ 5\ 6\ 7) = (1\ 2)(2\ 3)(4\ 5)(5\ 6)(6\ 7)$$

is an odd permutation whereas

$$(1\ 2\ 3\ 4\ 5)(6\ 7)(8\ 9\ a\ b) = (1\ 2)(2\ 3)(3\ 4)(4\ 5)(6\ 7)(8\ 9)(9\ a)(a\ b)$$

is an even permutation. It follows from Proposition 8.6 that every permutation is necessarily either even or odd, or both. However, it follows from Lemma 8.12 that

$$\sigma \Delta_n = \begin{cases} \Delta_n & \text{if } \sigma \text{ is even,} \\ -\Delta_n & \text{if } \sigma \text{ is odd.} \end{cases}$$

Since $\Delta_n \neq -\Delta_n$ (Exercise 8.4.27), we conclude that no permutation σ can be both odd and even. This is an important fact that deserves being stated as a proposition.

Proposition 8.14 A permutation σ of $\{1, 2, \dots, n\}$, $n \geq 2$, is either even or odd if $\sigma \Delta_n = \Delta_n$ or $-\Delta_n$, respectively. Consequently, no permutation can be both even and odd, and Δ_n has only two distinct variants.

An alternate proof of this proposition is indicated in Exercise 8.4.17. The *parity of a permutation* is its evenness or oddness.

Since it was already seen in the proof of Proposition 8.6 that every k -cycle can be expressed as the composition of $k - 1$ transpositions, the parity of a permutation can be easily computed from its disjoint cycle decomposition. Specifically, if $\sigma_1 \sigma_2 \cdots \sigma_m$ is the disjoint cycle decomposition of σ , where each σ_i is a k_i -cycle, then σ can be expressed as the composition of $\sum_{i=1}^m (k_i - 1)$ transpositions. Thus, the parity of σ is identical with the parity of the integer $\sum_{i=1}^m (k_i - 1)$. In particular, the parity of $(1\ 2\ 3\ 4)(5\ 6\ 7)$ is the same as the parity of $(4 - 1) + (3 - 1) = 5$ which is odd.

This notion of parity can be used to give another convincing example of the utility of permutations in proving negative results. The well-known *15-puzzle* consists of 15 square pieces, numbered 1 through 15, that are placed inside a larger square frame as indicated in Figure 8.1. A *legitimate move* consists of the sliding of a neighboring piece into the empty space. Figure 8.1 describes the effect of several successive legitimate moves. One now faces the challenge of rearranging the pieces into any prescribed configuration by means of legitimate moves alone. It turns out that some configurations, such as the

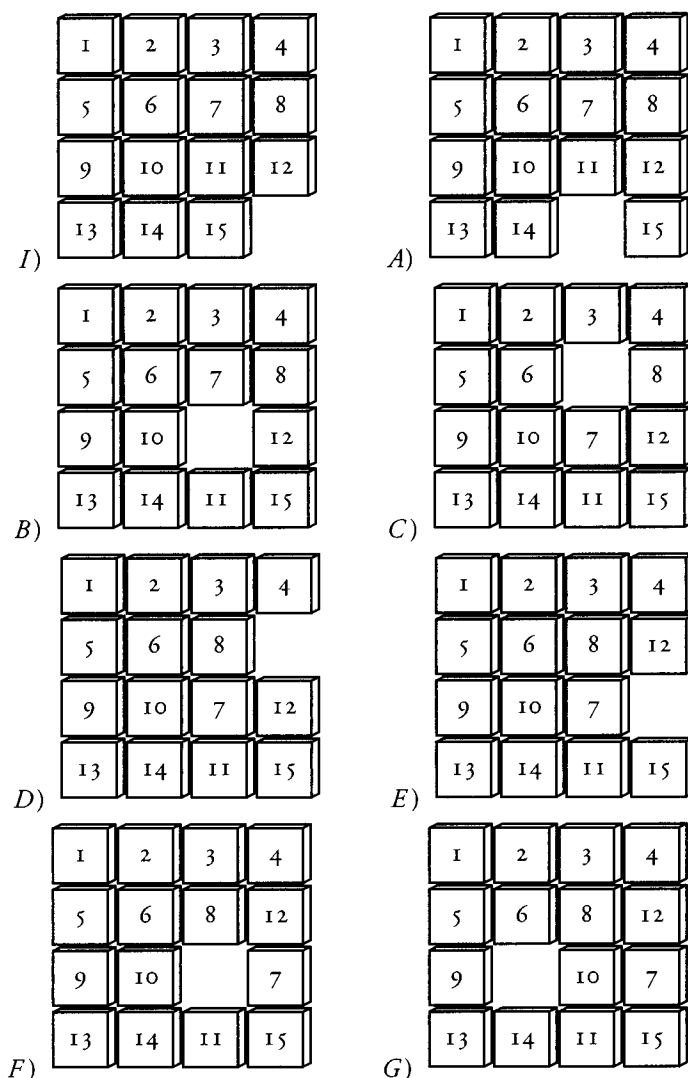


Figure 8.1 Legitimate moves for the 15-puzzle

configuration R (for reverse) of Figure 8.2, are in fact unattainable by means of legitimate moves. Parities of permutations can be used to give a rigorous proof of this fact, and we shall do so here.

Let I be the initial configuration of the 15-puzzle as described in Figure 8.1. To every prescribed configuration X of the 15-puzzle we assign a permutation P_X of the set $\{1, 2, 3, \dots, 15, b\}$ by

- (a) thinking of the empty space as just another piece labeled b (for blank); and
- (b) setting $P_X(i)$ to be the label of the piece that occupies in X the same location that i has in I .

For the configurations of Figure 8.1,

$$P_I = \text{Id},$$

$$P_A = (15\ b),$$

$$P_B = (15\ 11\ b),$$

$$P_C = (7\ b\ 15\ 11),$$

$$P_D = (7\ 8\ b\ 15\ 11),$$

$$P_E = (7\ 8\ 12\ b\ 15\ 11),$$

$$P_F = (7\ 8\ 1\ 2)(b\ 15\ 11),$$

$$P_G = (7\ 8\ 1\ 2)(10\ b\ 15\ 11).$$

Note that the sliding of some square labeled x , say, into the empty space, is tantamount to the transposition $(x\ b)$ and hence we have the following lemma.

Lemma 8.15 If the configuration Y is obtained from the configuration X by sliding the piece x into the empty space, then $P_Y = (x\ b)P_X$.

In Figure 8.1, F is obtained from E by sliding 7 into the empty space, and indeed

$$P_F = (7\ 8\ 1\ 2)(b\ 15\ 11) = (7\ b)(7\ 8\ 12\ b\ 15\ 11) = (7\ b)P_E.$$

It follows from this lemma that if a configuration X is obtained from the initial configuration I by a sequence of m moves, then P_X can be expressed as the composition of m transpositions.

We will now show that the configuration R of Figure 8.2 is unattainable by legitimate moves. Observe that

$$P_R = (1\ 15)(2\ 14)(3\ 13)(4\ 12)(5\ 11)(6\ 10)(7\ 9)$$

is an odd permutation, since it is expressed here as the composition of seven transpositions. On the other hand, since the empty space occupies the same positions in the initial configuration I and in R , it follows that a sequence of legitimate moves leading from I

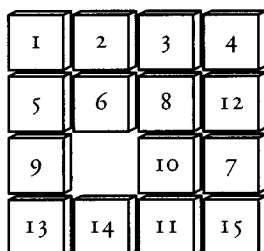


Figure 8.2 An unattainable configuration

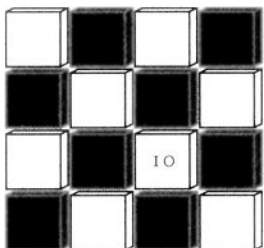


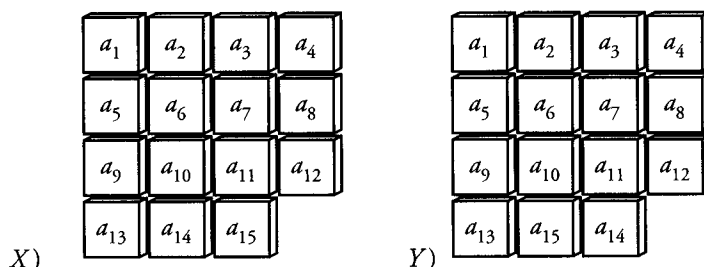
Figure 8.3 A coloring of the 15-puzzle

to R must consist of an even number of moves. One way to justify this assertion is to note that since such a sequence of moves terminates with b in its original location, the number of vertical moves must have been even, as must have been the number of horizontal moves. An alternate justification is obtained by coloring the underlying framework in the checkerboard pattern of Figure 8.3. Each legitimate move then changes the color showing in the empty space. Since the empty space returned to its original position, this sequence must consist of an even number of moves. Either way, Proposition 8.14 guarantees that the odd permutation P_R is not expressible by means of an even number of transpositions and hence the configuration R is not attainable by legitimate moves.

A configuration X is said to be *standard* if the empty space occupies the same position in X as it does in I , i.e., if $P_X(b) = b$. It is clear that the above argument proves the unattainability of any standard configuration X for which P_X is an odd permutation.

Proposition 8.16 A standard configuration X is attainable from the initial configuration I if and only if P_X is an even permutation.

Sketch of proof. If X is a standard configuration with P_X odd, then the argument that was applied to prove the unattainability of R above accomplishes the same goal for X .

Figure 8.4 Either X or Y is attainable

The proof of the converse is based on the observation (Exercise 8.4.29) that regardless of the values of the a_i 's, one of the two configurations of Figure 8.4 is necessarily attainable by legitimate moves. Moreover, the permutations P_X and P_Y associated with these configurations are related by $P_Y = (a_{14} a_{15})P_X$. Consequently, if P_X is even, then P_Y is odd and hence Y is unattainable. Since we have argued that either X or Y must be attainable, it follows that X must be attainable. ■

Exercises 8.4

Determine the parities of the permutations in Exercises 8.4.1 to 8.4.4.

o	even	odd
even	even	odd
odd	odd	even

Table 8.1 Parities of permutations

- | | |
|---|---|
| <p>1. $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 9 & 3 & 4 & 7 & 1 & 5 & 2 & 6 & 8 \end{pmatrix}$</p> <p>2. $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 5 & 3 & 4 & 7 & 1 & 6 & 2 & 9 & 8 \end{pmatrix}$</p> | <p>3. $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 9 & 8 & 7 & 6 & 5 & 4 & 3 & 2 & 1 \end{pmatrix}$</p> <p>4. $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 9 & 7 & 5 & 3 & 1 & 8 & 6 & 4 & 2 \end{pmatrix}$</p> |
|---|---|

5. List all the even permutations of $\{1, 2\}$ in disjoint cycle form.
6. List all the even permutations of $\{1, 2, 3\}$ in disjoint cycle form.
7. List all the even permutations of $\{1, 2, 3, 4\}$ in disjoint cycle form.
8. List all the odd permutations of $\{1, 2\}$ in disjoint cycle form.
9. List all the odd permutations of $\{1, 2, 3\}$ in disjoint cycle form.

10. List all the odd permutations of $\{1, 2, 3, 4\}$ in disjoint cycle form.
11. Prove that the effect of composition on the parities of permutations is described by Table 8.1.
12. Prove that if ρ and σ are any two permutations, then
 - (a) $\sigma\rho$ and $\rho\sigma$ have the same parities;
 - (b) σ and $\rho\sigma\rho^{-1}$ have the same parities.
13. Prove that every even permutation is expressible as the composition of 3-cycles and that this is not true for odd permutations.
14. Let $k > 1$ be an odd positive integer. Prove that every even permutation of $\{1, 2, \dots, n\}$, $n \geq k$, is expressible as the composition of k -cycles and that this is not true for odd permutations.
15. Prove that for $n \geq 3$ every even permutation of $\{1, 2, \dots, n\}$ is expressible as a composition of 3-cycles of the form $(1\ a\ b)$ and $(1\ 2\ a)$.
16. Determine for which n there exist cyclic permutations σ and τ of $\{1, 2, \dots, n\}$ such that $\sigma\tau = (1\ 2\ 3\ \dots\ n)$. Prove your answer.
17. Use Exercise 8.2.24 to prove Proposition 8.14.
18. Which of the configurations of the 15-puzzle in Figure 8.5 are attainable by legitimate moves?
19. Formulate and prove a necessary and sufficient condition for an arbitrary (not necessarily standard) configuration to be attainable by legitimate moves.
20. Which of the configurations of Figure 8.6 is attainable?
21. How many of the standard configurations of the 15-puzzle are attainable and how many are not? Prove your answer.
22. Prove that for $n > 1$, exactly half of the permutations of $\{1, 2, \dots, n\}$ are even.

For the values of n and k specified in Exercises 8.4.23 to 8.4.26, does there exist a function of n variables that has k variants? Justify your answers.

23. $n \geq 3, k = 2n$
24. $n \geq 2, k = 2\binom{n}{2}$
25. $n \geq 3, k = 2\binom{n}{3}$
26. $n > r + 1 \geq 2, k = 2\binom{n}{r}$
27. Let σ be a permutation of $\{1, 2, \dots, n\}$. Prove that the parity of the permutation σ is the same as the parity of the integer $n - \|\sigma\|$.
28. Prove that for $n \geq 2$, $\Delta_n \neq -\Delta_n$.

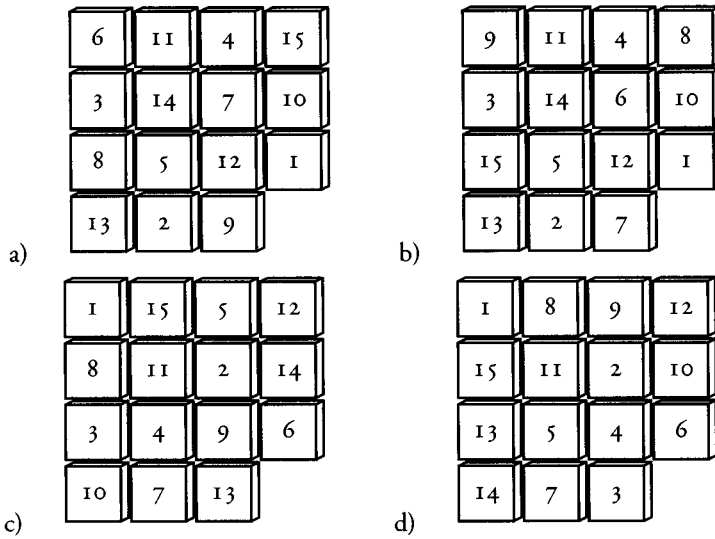


Figure 8.5 Some configurations of the 15-puzzle

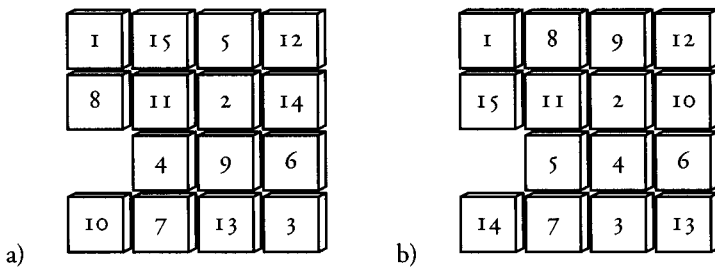


Figure 8.6 More configurations of the 15-puzzle

29. Complete the proof of Proposition 8.16 by showing that regardless of the values of the a_i 's, one of the two configurations of Figure 8.4 is necessarily attainable by legitimate moves.

Chapter Summary

Motivated by Lagrange's solution of the general quartic equation, we studied the number of distinct variants that a multivariable function can have. It was shown in Theorem 8.10 that there are some strong and surprising limitations on the number of such variants. The proof called for a deep analysis of the structure of permutations. The same proof also led to the formulation of the notion of the parity of a permutation. Finally, these

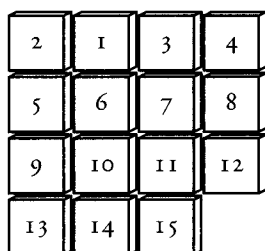


Figure 8.7 A configuration of the 15-puzzle

new tools, developed with functions in mind, were applied toward the resolution of the popular 15-puzzle.

Chapter Review Exercises

Mark the following true or false.

1. The number of distinct variants of the function $x_1(x_2 - x_3)^2$ is six.
2. The number of permutations of $\{a, b, c, d, e\}$ is 120.
3. $(1\ 2\ 3)(4\ 7\ 6\ 5)(1\ 7\ 6\ 2\ 3)(4\ 5) = (1\ 6\ 3\ 2)(5\ 7)$.
4. The permutation $(7\ 1\ 4\ 3\ 6\ 2)(5\ 9)(8)$ is expressible as the composition of transpositions.
5. If $f = x_1 + 2x_2 + 3x_3 + 4x_4 + 5x_5$, then $(1\ 2\ 3\ 4\ 5)f = (3\ 4\ 5\ 1\ 2)f$.
6. $(1\ 2\ 3)(2\ 3\ 4)f = (1\ 2)(3\ 4)f$.
7. There is a function of eight variables which has four variants.
8. There is a function of eight variables which has two variants.
9. The 15-puzzle configuration of Figure 8.7 is attainable by legitimate moves.

New Terms

15-puzzle, 171

cycle, 161

cyclic permutation, 161

discriminant, 169

disjoint cycle decomposition, 162

distinct variants, 156

even permutation, 170

identity permutation, 158

invariant function, 156

odd permutation, 170

parity of a permutation, 171

permutation, 158

transposition, 163

variants, 155

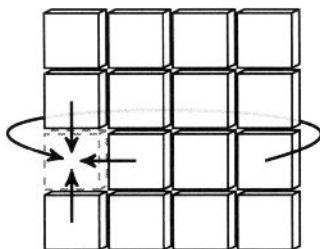


Figure 8.8 Legitimate moves for the cylindrical $(2, 2)$ -puzzle

Supplementary Exercises

If k and n are any two positive integers, the (k, n) -puzzle is similar to the 15-puzzle, except that it consists of a rectangular array of squares containing k rows and n columns. Accordingly, the $(4, 4)$ -puzzle is identical with the 15-puzzle analyzed in this chapter.

1. Given any initial configuration of the $(4, 5)$ -puzzle, describe all the configurations attainable from it.
2. Given any initial configuration of the $(5, 5)$ -puzzle, describe all the configurations attainable from it.
3. For any positive integers k and n , given an initial configuration of the (k, n) -puzzle, describe all the configurations attainable from it.
4. The *cylindrical* (k, n) -puzzle is obtained from the traditional version by adding some moves that allow for the pieces to cross the vertical boundaries of the board. Figure 8.8 illustrates some legitimate moves for the cylindrical version of the puzzle. The reason for the *cylindrical* appellation is that it is possible to obtain a physical interpretation of the legitimacy of the new moves by bending the puzzle into a cylinder and gluing its vertical edges so that in this new form the pieces always do slide into adjacent spaces. Given any initial configuration of the cylindrical (k, n) -puzzle, decide which configurations are attainable from it.
5. The *toroidal* (k, n) -puzzle is obtained from the cylindrical puzzle by allowing the pieces to slide into the empty space across the horizontal boundaries as well (Figure 8.9). Given any initial configuration of the toroidal (k, n) -puzzle, decide which configurations are attainable from it. This game can be visualized as taking place on a torus (Figure 8.10).

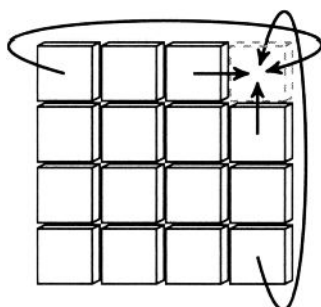


Figure 8.9 Legitimate moves for the toroidal $(2, 2)$ -puzzle



Figure 8.10 A torus

6. The *Möbius* (k, n) -puzzle is obtained from the traditional one by allowing for the filling of the blanks across the vertical boundaries, with the additional twist that the blank can be filled only by boundary pieces that are diametrically opposite to it (Figure 8.11). As its name implies, this version can be visualized on the Möbius strip of Figure 8.12. For any initial configuration of the Möbius (k, n) -puzzle, determine the configurations that are attainable from it.
7. The *Klein bottle* (k, n) -puzzle allows for the cylindrical moves of Exercise 8.s.4 across its horizontal boundaries and the Möbius-type moves of Exercise 8.s.6 across its vertical boundaries. For any initial configuration of the Klein bottle (k, n) -puzzle, determine the configurations that are attainable from it. This game can be visualized (with some difficulty) on the Klein bottle of Figure 8.13.

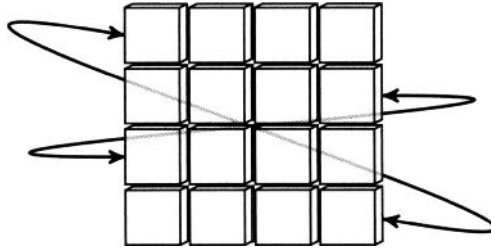


Figure 8.11 Examples of diametrically opposite boundary pieces

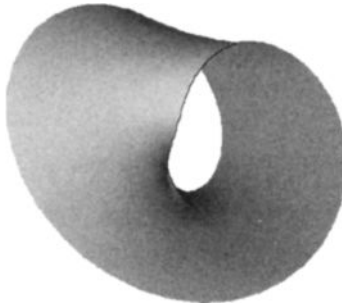


Figure 8.12 The Möbius strip

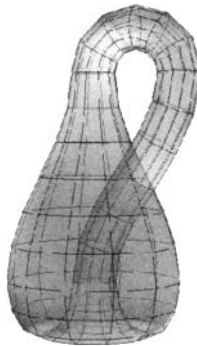


Figure 8.13 The Klein bottle

8. Generalize the (k, n) -puzzle to other planar versions.
9. Formulate a 3-dimensional version of the 15-puzzle and solve it.
10. Formulate a 3-dimensional version of the toroidal (k, n) -puzzle and solve it.
11. Formulate a 3-dimensional version of the Klein bottle (k, n) -puzzle and solve it.
12. Formulate and solve d -dimensional analogs of all the versions of the (k, n) -puzzle for every positive integer d .
13. Prove Cauchy's Theorem that a function of $n \geq 5$ variables which has fewer than n distinct variants must have either one or two distinct variants.
14. Let n and k be two positive integers. Which permutations of $\{1, 2, \dots, n\}$ are expressible as the composition of k cyclic permutations of the same set?
15. Is it true that for any three integers $k, l, m > 1$ there exist permutations ρ and σ such that ρ , σ , and $\rho\sigma$ have orders k , l , and m , respectively?

Chapter 9



GROUPS

WE NOW EXTRACT the notion of a group from the variety of algebraic structures that have been discussed in the previous chapters. This leads to a natural classification problem and information is provided toward the classification of some elementary groups.

9.1 Permutation Groups

A few years after Abel proved that the general quintic equation was not solvable by radicals the young Évariste Galois discovered a general criterion for determining whether any given equation was solvable by radicals. The only two specific examples that Galois gave were Abel's aforementioned theorem and Gauss's work on the cyclotomic equation, which was discussed in some detail in Chapters 2 and 3. Galois observed that permutations (he sometimes called them substitutions) played a crucial role in both proofs. As was seen in Chapter 8, these permutations are quite explicit in Abel's work. They are less evident in Gauss's proof, and the best that can be said within the confines of this book is that in the special case discussed in Section 2.4, the crucial permutation is

$$\sigma = (\zeta \zeta^3 \zeta^9 \zeta^{10} \zeta^{13} \zeta^5 \zeta^{15} \zeta^{11} \zeta^{16} \zeta^{14} \zeta^8 \zeta^7 \zeta^4 \zeta^{12} \zeta^2 \zeta^6)$$

where ζ is any primitive 17-th root of 1. Of course, σ can also be thought of as the algebraically defined function $\sigma(\alpha) = \alpha^3$ defined on the imaginary 17-th roots of 1, but we shall regard it as a purely formal permutation. This permutation bears an obvious relation to the sum

$$\zeta + \zeta^3 + \zeta^9 + \zeta^{10} + \zeta^{13} + \zeta^5 + \zeta^{15} + \zeta^{11} + \zeta^{16} + \zeta^{14} + \zeta^8 + \zeta^7 + \zeta^4 + \zeta^{12} + \zeta^2 + \zeta^6$$

that was used to prove the ruler-and-compass constructibility of the regular 17-sided polygon. The work of Lagrange, Gauss, and Abel eventually led Galois to formulate

the notion of a group of permutations, or permutation group. Given a set S , a *group of permutations* of S is a set G of permutations of S such that

- (a) G contains the identity permutation,
- (b) if σ is in G , so is σ^{-1} ,
- (c) if ρ and σ are in G , so is their composition $\rho\sigma$.

We have already encountered several groups of permutations. For each positive integer n , the group of all the permutations of the set $\{1, 2, \dots, n\}$ is called the *symmetric group* and is denoted by S_n . That S_n satisfies the three requirements above follows from the fact that it consists of all of the permutations of $\{1, 2, \dots, n\}$.

The list of permutations $\text{Id}, (1\ 2), (3\ 4), (1\ 2)(3\ 4)$, which consists of all the permutations that leave the polynomial $x_1 + x_2 - x_3 - x_4$ unchanged, is also a group of permutations. To see this it need only be observed that each of these permutations is its own inverse, and that the composition of any two distinct nonidentity elements is equal to the third nonidentity element. This is no coincidence.

Proposition 9.1 If f is any function of x_1, x_2, \dots, x_n , then the set $S_{n,f}$ of all the permutations of these variables that leave f unchanged is a group of permutations.

Proof. This follows from Lemma 8.8 and Corollary 8.9. ■

This proposition provides us with a host of groups of permutations. Thus, if $f = x_1 + x_2 + x_3$, then $S_{3,f} = S_3$. If $f = x_1 x_2^2 x_3^3$, then $S_{3,f} = \{\text{Id}\}$. For $f = x_1 x_2 + x_3$, $S_{3,f} = \{\text{Id}, (1\ 2)\}$, and finally, for

$$f = x_1 x_2^2 x_3^3 + x_1^2 x_2 x_3^3 + x_1^2 x_2^2 x_3$$

we have $S_{3,f} = \{\text{Id}, (1\ 2\ 3), (1\ 3\ 2)\}$. The converse of Proposition 9.1 also holds.

Proposition 9.2 Let G be any group of permutations of $\{1, 2, \dots, n\}$. Then there is a polynomial function f such that $G = S_{n,f}$.

Proof. It is clear that the polynomial $g = x_1 x_2^2 x_3^3 \cdots x_n^n$ is such that $S_{n,g} = \{\text{Id}\}$. Suppose now that $G = \{\sigma_1, \sigma_2, \dots, \sigma_k\}$ is a listing of the elements of G . Set

$$f = \sigma_1 g + \sigma_2 g + \cdots + \sigma_k g.$$

Since G is a group, it follows that for every $\sigma, \sigma_i \in G$, $\sigma\sigma_i$ also belongs to G . Moreover, $\sigma\sigma_i = \sigma\sigma_j$ if and only if $\sigma_i = \sigma_j$, if and only if $i = j$, so that $\sigma\sigma_1, \sigma\sigma_2, \dots, \sigma\sigma_k$ also

constitutes a listing of the elements of G . Hence, for any $\sigma \in G$,

$$\sigma f = \sigma \sigma_1 g + \sigma \sigma_2 g + \cdots + \sigma \sigma_k g = \sigma_1 g + \sigma_2 g + \cdots + \sigma_k g = f$$

so that $\sigma \in S_{n,f}$. Thus we have shown that $S_{n,f} \supset G$.

Conversely, suppose $\sigma \in S_{n,f}$. Then σ must transform each summand of f into another summand of f . In particular, for some $j = 1, 2, \dots, k$,

$$\sigma \sigma_1 g = \sigma(\sigma_1 g) = \sigma_j g \quad \text{or} \quad \sigma_j^{-1} \sigma \sigma_1 g = g.$$

Since $S_{n,g} = \{\text{Id}\}$, it follows that $\sigma_j^{-1} \sigma \sigma_1 = \text{Id}$ and so $\sigma = \sigma_j \text{Id} \sigma_1^{-1} = \sigma_j \sigma_1^{-1} \in G$. Thus, $S_{n,f} = G$. ■

If $G = \{\text{Id}, (1\ 2\ 3), (1\ 3\ 2)\}$, then $g = x_1 x_2^2 x_3^3$ and

$$f = x_1 x_2^2 x_3^3 + x_2 x_3^2 x_1^3 + x_3 x_1^2 x_2^3.$$

The group of permutations G that Galois associated with the cyclotomic equation $x^{17} - 1 = 0$ consists of all the powers of the permutation

$$\sigma = (\zeta \zeta^3 \zeta^9 \zeta^{10} \zeta^{13} \zeta^5 \zeta^{15} \zeta^{11} \zeta^{16} \zeta^{14} \zeta^8 \zeta^7 \zeta^4 \zeta^{12} \zeta^2 \zeta^6).$$

Put differently $G = \{\text{Id}, \sigma, \sigma^2, \sigma^3, \dots, \sigma^{15}\}$, where

$$\sigma^2 = (\zeta \zeta^9 \zeta^{13} \zeta^{15} \zeta^{16} \zeta^8 \zeta^4 \zeta^2)(\zeta^3 \zeta^{10} \zeta^5 \zeta^{11} \zeta^{14} \zeta^7 \zeta^{12} \zeta^6)$$

$$\vdots$$

$$\sigma^{16} = \text{Id}.$$

This generalizes as follows.

Proposition 9.3 Let σ be any permutation of $\{1, 2, \dots, n\}$, and let d be the order of σ . Then the set $\langle \sigma \rangle = \{\text{Id}, \sigma, \sigma^2, \dots, \sigma^{d-1}\}$ is a group of permutations.

Proof. The identity permutation Id belongs to $\langle \sigma \rangle$ by definition. Since $\sigma^d = \text{Id}$, it follows that $(\sigma^k)^{-1} = \sigma^{-k} = \sigma^{d-k}$ for $k = 0, 1, 2, \dots, d-1$ and so every element of $\langle \sigma \rangle$ has its inverse in $\langle \sigma \rangle$. Finally, if σ^k and σ^m are two arbitrary elements of $\langle \sigma \rangle$, then so is $\sigma^k \sigma^m = \sigma^{k+m}$ an element of $\langle \sigma \rangle$. Thus, $\langle \sigma \rangle$ is indeed a group of permutations. ■

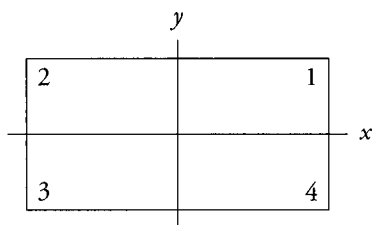


Figure 9.1 The symmetries of the rectangle

Accordingly,

$$\langle (1\ 2\ 3)(4\ 5) \rangle = \{ \text{Id}, (1\ 2\ 3)(4\ 5), (1\ 3\ 2), (4\ 5), (1\ 2\ 3), (1\ 3\ 2)(4\ 5) \}$$

and

$$\begin{aligned} \langle (1\ 2\ 3\ 4\ 5\ 6) \rangle = & \{ \text{Id}, (1\ 2\ 3\ 4\ 5\ 6), (1\ 3\ 5)(2\ 4\ 6), (1\ 4)(2\ 5)(3\ 6) \} \\ & \cup \{ (1\ 5\ 3)(2\ 6\ 4), (1\ 6\ 5\ 4\ 3\ 2) \} \end{aligned}$$

are groups of permutations. The *alternating group* A_n consists of the set of all the even permutations of n symbols. Thus,

$$A_1 = \{ \text{Id} \}, \quad A_2 = \{ \text{Id} \}, \quad A_3 = \{ \text{Id}, (1\ 2\ 3), (1\ 3\ 2) \},$$

and

$$\begin{aligned} A_4 = & \{ \text{Id}, (1\ 2)(3\ 4), (1\ 3)(2\ 4), (1\ 4)(2\ 3), (1\ 2\ 3), (1\ 3\ 2) \} \\ & \cup \{ (1\ 2\ 4), (1\ 4\ 2), (1\ 3\ 4), (1\ 4\ 3), (2\ 3\ 4), (2\ 4\ 3) \}. \end{aligned}$$

The formal proof of the fact that A_n is indeed a group of permutations is relegated to Exercise 9.1.17.

There is a host of groups of permutations that are defined geometrically in terms of symmetries of configurations rather than algebraic symbols. Consider the rectangle of Figure 9.1 whose vertices are labeled 1, 2, 3, 4. It has two obvious symmetries, with respect to the x - and y -axes. The first of these interchanges the vertices 1 and 4 and also 2 and 3. Thus it can be denoted by the permutation $(1\ 4)(2\ 3)$. Similarly, the symmetry with

respect to the y -axis induces the permutation $(1\ 2)(3\ 4)$ on the vertices. The central symmetry that the rectangle has with respect to the origin induces the permutation $(1\ 3)(2\ 4)$ on the vertices. Each of these three permutations is its own inverse, and the composition of any two distinct ones equals the third. It therefore follows that if we add Id as a trivial symmetry, then the set

$$K = \{ \text{Id}, (1\ 2)(3\ 4), (1\ 3)(2\ 4), (1\ 4)(2\ 3) \}$$

is a group of permutations. This group is known as the *Klein 4-group*, after the mathematician Felix Klein whose *Erlanger Programm* of 1872 set the tone for the investigation of the relationship between geometry and group theory for generations to come.

Before we go on to describe some more geometrical groups of permutations, it is necessary to firm up the notion of a symmetry. For the purposes of this discussion, the geometrical configurations in question are all assumed to be centered at the origin of a three-dimensional coordinate system, and a symmetry of the configuration is a rotation of the ambient space about an axis through the origin that leaves the position of the configuration unchanged. Thus, from our point of view, a symmetry with respect to the x -axis results from a 180° rotation of space about the x -axis. This rotation clearly transforms the rectangle of Figure 9.1 right back onto itself. The central symmetry of the rectangle with respect to the origin comes from a 180° rotation of space about the z -axis which is not drawn in the figure. It is clear that by this definition every configuration possesses at least one symmetry, namely, the trivial one that is defined by the identity transformation of space. Since each symmetry transforms the configuration onto itself, each such symmetry must necessarily permute the vertices of the configuration, as was the case in the above rectangle. We refer to these permutations as *vertex symmetries*.

Proposition 9.4 The vertex symmetries of any geometrical configuration form a group of permutations.

Proof. As was noted above, the identity permutation is a vertex symmetry of any configuration. Since the inverse of any rotation is also a rotation, it follows that the inverse of any vertex symmetry is also a vertex symmetry. Finally, since the composition of any two rotations whose axes intersect at the origin is known to be another such rotation it follows that the composition of any two vertex symmetries is also a vertex symmetry. ■

We emphasize here that if σ is the vertex permutation that describes the rotation R , then the rotation R replaces the vertex v with the vertex $\sigma(v)$. The square of Figure 9.2

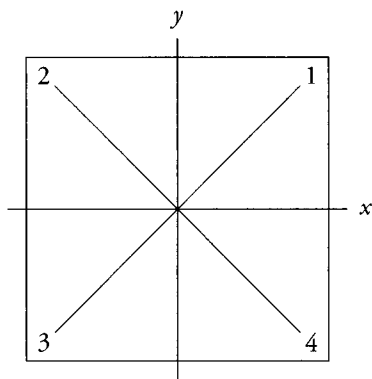


Figure 9.2 The symmetries of the square

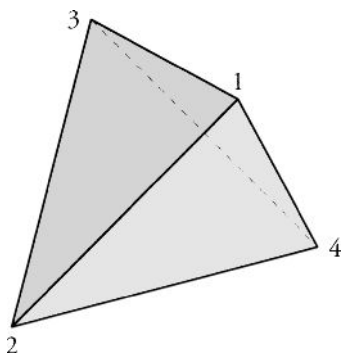


Figure 9.3 A regular tetrahedron

has eight symmetries: the four that it has as a rectangle, two more that result from clockwise and counterclockwise 90° rotations about the z -axis, and two more that result from 180° rotations about the diagonals 13 and 24, respectively. These last four induce the following respective vertex permutations: $(1\ 2\ 3\ 4)$, $(1\ 4\ 3\ 2)$, $(2\ 4)$, and $(1\ 3)$. Thus, the vertex symmetries of the square constitute the permutation group

$$D_4 = \{ \text{Id}, (1\ 2)(3\ 4), (1\ 3)(2\ 4), (1\ 4)(2\ 3), (1\ 2\ 3\ 4), (1\ 4\ 3\ 2), (1\ 3), (2\ 4) \}.$$

In general, the group of vertex symmetries of the regular n -gon is called the *dihedral group* D_n .

We turn next to some interesting vertex symmetry groups defined by solid configurations. The regular tetrahedron of Figure 9.3, with vertices 1, 2, 3, and 4 has four faces

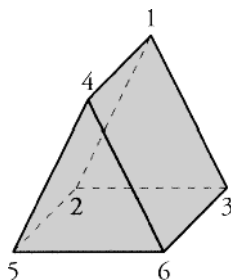


Figure 9.4 A triangular prism.

each of which is an equilateral triangle. Each altitude (the line joining a vertex to the center of the opposite face) serves as the axis of two nontrivial $\pm 120^\circ$ rotations, thus contributing a total of eight vertex symmetries:

$$\begin{array}{cccc} (1\ 2\ 3), & (1\ 3\ 2), & (1\ 2\ 4), & (1\ 4\ 2), \\ (1\ 3\ 4), & (1\ 4\ 3), & (2\ 3\ 4), & (2\ 4\ 3). \end{array}$$

In addition, the line joining the midpoints of the two edges 23 and 14 serves as the axis of a 180° rotation that defines the vertex symmetry $(2\ 3)(1\ 4)$ with the analogous lines defining the additional vertex symmetries $(1\ 3)(2\ 4)$ and $(1\ 2)(3\ 4)$.

It is now clear that the vertex symmetries of the tetrahedron constitute a by now familiar group, namely, the alternating group A_4 that consists of all the even permutations of $\{1, 2, 3, 4\}$.

This is the time to note that we have excluded some symmetries that others might, and sometimes do, include. Specifically, it could be argued that the tetrahedron of Figure 9.3 possesses a symmetry with respect to the plane that contains the edge 14 and bisects the edge 23. The decision to restrict our attention only to symmetries that are realizable as rotations in three-dimensional space was arbitrary and based on pedagogical grounds.

Consider the triangular prism of Figure 9.4, whose lateral sides are equilateral triangles. Its group of vertex symmetries consists of the elements

$$\begin{array}{ccc} \text{Id}, & (1\ 2\ 3)(4\ 5\ 6), & (1\ 3\ 2)(4\ 6\ 5), \\ (1\ 5)(2\ 4)(3\ 6), & (1\ 4)(2\ 6)(3\ 5), & (1\ 6)(2\ 5)(3\ 4). \end{array}$$

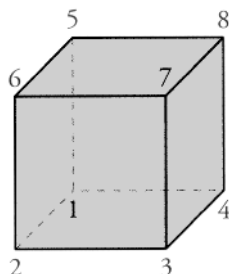


Figure 9.5 A cube

Exercises 9.1

Which of the sets of permutations of $\{1, 2, 3, 4, 5\}$ in Exercises 9.1.1 to 9.1.6 form a group? Justify your answer.

1. All the even permutations.
2. All the odd permutations.
3. All the transpositions.
4. All the permutations that leave 3 fixed.
5. All the permutations that interchange 2 and 4.
6. All the permutations that map 1 to 5.

List the elements of the group $S_{n,f}$ for the function f in Exercises 9.1.7 to 9.1.16.

- | | |
|-------------------------------|----------------------------------|
| 7. $x_1 + x_2 + 2x_3$ | 12. $(x_1 + x_2)x_3x_4x_5$ |
| 8. $x_1x_2 + x_2x_3 + x_3x_1$ | 13. $(x_1 - x_2)(x_3 - x_4)$ |
| 9. $x_1 + x_2 + x_3 - x_4$ | 14. $(x_1 + x_2 - x_3 - x_4)x_5$ |
| 10. $(x_1 + x_2)(x_3 + x_4)$ | 15. $(x_1 + x_2 - x_3 - x_4)^2$ |
| 11. $x_1(x_2 + x_3 - x_4)$ | 16. $x_1/x_2 + x_3/x_4$ |

17. Prove that if $f = \Delta_n$ then $S_{n,f} = A_n$.
18. Describe the vertex symmetries of the cube of Figure 9.5.
19. Describe the vertex symmetries of the octahedron of Figure 9.6.
20. Describe, without listing them, all the vertex symmetries of the dodecahedron of Figure 9.7.

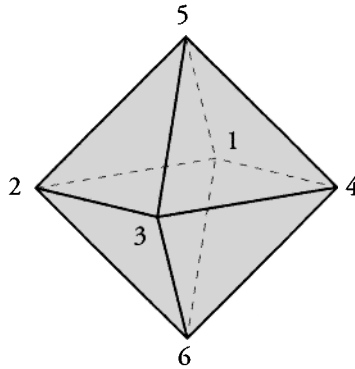


Figure 9.6 A regular octahedron

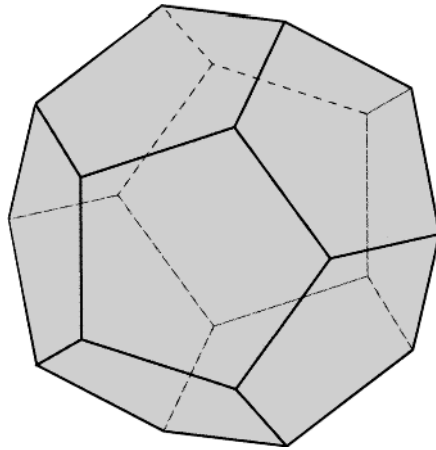


Figure 9.7 A regular dodecahedron

21. Describe, without listing them, all the vertex symmetries of the icosahedron of Figure 9.8.

Let $A = (1\ 2\ 3)$, $B = (2\ 4\ 3)$, $C = (1\ 2)(3\ 4)$, and $D = (1\ 3)(2\ 4)$ be four rotations of the tetrahedron of Figure 9.3. Find the axes and the angle of the rotations in Exercises 9.1.22 to 9.1.27.

22. $A \circ B$

24. $B \circ A$

26. $C \circ A$

23. $A \circ C$

25. $C \circ B$

27. $D \circ C$

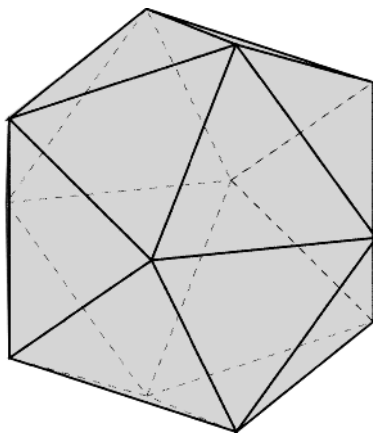


Figure 9.8 A regular icosahedron

Let $A = (1\ 2\ 3\ 4)(5\ 6\ 7\ 8)$, $B = (1\ 8\ 6)(2\ 4\ 7)$, and $C = (1\ 2)(7\ 8)(4\ 6)(3\ 5)$ be three rotations of the cube of Figure 9.5. Find the axes and the angle of the rotations in Exercises 9.1.28 to 9.1.33.

28. $A \circ B$

30. $B \circ A$

32. $C \circ A$

29. $A \circ C$

31. $C \circ B$

33. $B \circ C$

34. List the elements of the dihedral group D_5 as permutations.
35. How many elements does the dihedral group D_n have?
36. How many of the elements of the dihedral group D_n have order 2?
37. Prove that for every positive integer k there is a polygon whose group of vertex symmetries contains k elements.

9.2 Abstract Groups

Every group of permutations resembles the earlier algebraic structures of this text, such as the real numbers, the complex numbers, \mathbb{Z}_p , and the Galois fields, in that it involves a binary operation on a set of objects, this operation being, of course, the composition of permutations. Thus, it is possible to associate with every group of permutations a multiplication table that greatly resembles the tables that were associated with \mathbb{Z}_n for each positive integer n .

Table 9.1 describes the compositions of the elements of the Klein 4-group K . There, as elsewhere, the product ab is to be found at the intersection of row a and column b .

	Id	$(1\ 2)(3\ 4)$	$(1\ 3)(2\ 4)$	$(1\ 4)(2\ 3)$
Id	Id	$(1\ 2)(3\ 4)$	$(1\ 3)(2\ 4)$	$(1\ 4)(2\ 3)$
$(1\ 2)(3\ 4)$	$(1\ 2)(3\ 4)$	Id	$(1\ 4)(2\ 3)$	$(1\ 3)(2\ 4)$
$(1\ 3)(2\ 4)$	$(1\ 3)(2\ 4)$	$(1\ 4)(2\ 3)$	Id	$(1\ 2)(3\ 4)$
$(1\ 4)(2\ 3)$	$(1\ 4)(2\ 3)$	$(1\ 3)(2\ 4)$	$(1\ 2)(3\ 4)$	Id

 Table 9.1 The multiplication table of the Klein 4-group K

Similarly, the group of vertex symmetries of the square has Table 9.2 associated with it. It was the British mathematician Cayley who eventually extracted the notion of an abstract group from these tables (see Appendix E). Accordingly, an *abstract group* consists of a set G and a binary operation \cdot on its elements such that the following four properties are satisfied:

- for any two elements a and b of G , $a \cdot b$ is also in G ,
- there is an element 1_G of G such that $a \cdot 1_G = 1_G \cdot a = a$ for every element a of G ,
- for any elements $a, b, c \in G$, $(a \cdot b) \cdot c = a \cdot (b \cdot c)$,
- for every element a of G there is an element $a^\#$ such that $a \cdot a^\# = a^\# \cdot a = 1_G$.

Such an abstract group is denoted by either (G, \cdot) , or sometimes by G alone, if the binary operation is understood. The element 1_G is called the identity of G , and the element $a^\#$ is called the inverse of a .

It is clear that every group of permutations is an abstract group, with composition as the binary operation in question. The identity permutation Id functions as the identity 1_G and σ^{-1} functions as $\sigma^\#$ for every permutation σ . It is also clear that the set of the real numbers, together with the operation of addition, constitutes a group. Here 0 functions as the group identity and $a^\# = -a$ for all real a . This group will be denoted by $(\mathbb{R}, +)$. Similarly $(\mathbb{Z}, +)$, $(\mathbb{Q}, +)$, and $(\mathbb{C}, +)$ denote groups whose underlying sets are the integers, the rationals, and the complex numbers, respectively. On the other hand, the integers under multiplication do not constitute a group, since very few integers have multiplicative inverses. Nor do the rational, the real, or the complex numbers constitute a group under multiplication, since in each case 0 fails to have a multiplicative inverse.

We have already encountered many other groups in this book. Recalling that $\sqrt[n]{1}$ denotes the set of all the n complex roots of 1, we note that $(\sqrt[n]{1}, \cdot)$ is a group wherein \cdot denotes the ordinary multiplication of complex numbers. The identity element of this

	Id	(1 2 3 4)	(1 3)(2 4)	(1 4 3 2)	(1 3)	(2 4)	(1 2)(3 4)	(1 4)(2 3)
Id	Id	(1 2 3 4)	(1 3)(2 4)	(1 4 3 2)	(1 3)	(2 4)	(1 2)(3 4)	(1 4)(2 3)
(1 2 3 4)	(1 2 3 4)	(1 3)(2 4)	(1 4 3 2)	Id	(1 4)(2 3)	(1 2)(3 4)	(1 3)	(2 4)
(1 3)(2 4)	(1 3)(2 4)	(1 4 3 2)	Id	(1 2 3 4)	(2 4)	(1 3)	(1 4)(2 3)	(1 2)(3 4)
(1 4 3 2)	(1 4 3 2)	Id	(1 2 3 4)	(1 3)(2 4)	(1 2)(3 4)	(1 4)(2 3)	(2 4)	(1 3)
(1 3)	(1 3)	(1 2)(3 4)	(2 4)	(1 4)(2 3)	Id	(1 3)(2 4)	(1 2 3 4)	(1 4 3 2)
(2 4)	(2 4)	(1 4)(2 3)	(1 3)	(1 2)(3 4)	(1 3)(2 4)	Id	(1 4 3 2)	(1 2 3 4)
(1 2)(3 4)	(1 2)(3 4)	(2 4)	(1 4)(2 3)	(1 3)	(1 4 3 2)	(1 2 3 4)	Id	(1 3)(2 4)
(1 4)(2 3)	(1 4)(2 3)	(1 3)	(1 2)(3 4)	(2 4)	(1 2 3 4)	(1 4 3 2)	(1 3)(2 4)	Id

Table 9.2 The multiplication table of D_4

group is 1, the inverse ζ^\sharp of the root ζ is simply ζ^{-1} , also an element of $\sqrt[n]{1}$, and it is clear that if ζ and η are any two elements of $\sqrt[n]{1}$, then so is their product $\zeta\eta$ since $(\zeta\eta)^n = \zeta^n\eta^n = 1$.

For any integer n , addition modulo n defines a group $(\mathbb{Z}_n, +)$, with 0 acting as the group identity and $a^\sharp = -a$. However, when p is a prime integer, modular arithmetic can be used to define another, less expected, collection of groups. If p is a prime number, set $\mathbb{Z}_p^* = \{1, 2, 3, \dots, p-1\}$. Then we know that each element of \mathbb{Z}_p^* has a multiplicative inverse in \mathbb{Z}_p . Consequently, (\mathbb{Z}_p^*, \cdot) is a group with identity 1, where \cdot denotes multiplication modulo p . This collection of groups can be considerably enlarged as follows. For each positive integer n let \mathbb{Z}_n^* denote the set of positive integers that are both smaller than n and relatively prime to it. For example, $\mathbb{Z}_6^* = \{1, 5\}$, $\mathbb{Z}_8^* = \{1, 3, 5, 7\}$, and $\mathbb{Z}_{10}^* = \{1, 3, 7, 9\}$. Corollary 4.4 and Lemma 4.6 guarantee that (\mathbb{Z}_n^*, \cdot) is indeed a group, where \cdot is multiplication modulo n .

If F is any field, then $(F, +)$ also forms a group. Note that if $F = \text{GF}(2, P(x))$, then $a^\sharp = a$ for each a in F . Again, if F^* denotes all the nonzero elements of F , then (F^*, \cdot) is also a group, where \cdot denotes the multiplication operation in the field.

If F is any field, then $F[x]$, the set of polynomials with coefficients in F , is a group with respect to addition. Similarly, if n is any positive integer and $F[x, \leq n]$ denotes all the polynomials in $F[x]$ that have degree at most n together with the zero polynomial, then $F[x, \leq n]$ is also a group with respect to addition. For example, $\mathbb{Z}_2[x, \leq 1]$ has as elements 0, 1, x , and $1+x$.

Some groups can be defined directly by means of a multiplication table. The group whose multiplication table is Table 9.3 is called the *Quaternion group*. It is clear that in this multiplication table 1 functions as an identity and $a^\sharp = e$, $b^\sharp = f$, etc. The direct verification of the associativity of this multiplication table calls for several hundred computations. More efficient techniques are available, but they fall outside the confines of this text.

A group is said to be a *commutative group* (or abelian group) if for any two of its elements a and b we have $ab = ba$. Thus, the groups $(\mathbb{Z}_n, +)$, $\text{GF}(p, P(x))$, \mathbb{Z}_n^* , \mathbb{Z} , and \mathbb{C} are all commutative. On the other hand, if $n \geq 3$, then the group S_n is not commutative since in each such group

$$(1\ 2)(2\ 3) = (1\ 2\ 3) \neq (3\ 2\ 1) = (2\ 3)(1\ 2).$$

	1	a	b	c	d	e	f	g
1	1	a	b	c	d	e	f	g
a	a	d	c	f	e	1	g	b
b	b	g	d	a	f	c	1	e
c	c	b	e	d	g	f	a	1
d	d	e	f	g	1	a	b	c
e	e	1	g	b	a	d	c	f
f	f	c	1	e	b	g	d	a
g	g	f	a	1	c	b	e	d

Table 9.3 The multiplication table of the Quaternion group

A digression on the nature of multiplication tables might be in order here. It is clear that (the interior of) the multiplication table of a group with n elements consists of an n -by- n array. It is customary to list the rows and the columns of the array in the same order, with the row and column that correspond to the identity element appearing first. Each row and each column of the multiplication table constitutes a permutation of the elements of the group. A table that possesses all these properties is called a *Latin square*. Thus, the multiplication table of every group is a Latin square. The converse is not true. Most Latin squares do not come from groups, and Exercises 9.2.31 to 9.2.35 contain additional information on this subject.

The definition of an abstract group stipulates the existence of inverses, but says nothing about the possible existence of multiple inverses. The next proposition shuts the door on this possibility.

Proposition 9.5 If G is a group and $a \in G$, then a has exactly one inverse in G .

Proof. Suppose both b and c are inverses of a in G , i.e., $ba = ab = 1_G = ca = ac$. Then $b = b1_G = b(ac) = (ba)c = 1_Gc = c$. ■

The next proposition about inverses will prove useful later.

Proposition 9.6 If a and b are elements of the group G , then $(ab)^\# = b^\#a^\#$.

Proof. Several applications of the Associative Law yield

$$(b^\#a^\#)(ab) = b^\#(a^\#a)b = b^\#1_Gb = b^\#b = 1_G$$

and

$$(ab)(b^\sharp a^\sharp) = a(bb^\sharp)a^\sharp = a1_G a^\sharp = aa^\sharp = 1_G.$$

Thus, $b^\sharp a^\sharp$ acts like an inverse of ab and hence, by Proposition 9.5, it must be the inverse of ab , i.e., $b^\sharp a^\sharp = (ab)^\sharp$. ■

If a is any element of a group G , and n is any positive integer, then we define a^n as the product of n a 's. Thus, $a^1 = a$, $a^2 = aa$, $a^3 = aaa$, and so on. If we also define $a^0 = 1_G$ and $a^{-n} = (a^\sharp)^n$, then it is easily verified (Exercise 9.2.36) that, just like the powers of real numbers, the powers of abstract group elements satisfy the conditions $a^m a^n = a^{m+n}$ and $(a^m)^n = a^{mn}$ for any two integers m and n .

Exercises 9.2

Each of Exercises 9.2.1 to 9.2.14 specifies a set and a binary operation. In which cases do these form a group? If not, explain why not.

1. All the even elements of $\mathbb{Z}_{1,000}$ under addition
2. All the odd elements of $\mathbb{Z}_{1,000}$ under addition
3. All the even elements of $\mathbb{Z}_{1,000}$ under multiplication
4. All the odd elements of $\mathbb{Z}_{1,000}$ under multiplication
5. All the even elements of \mathbb{Z}_{64} under multiplication
6. All the odd elements of \mathbb{Z}_{64} under multiplication
7. All the integers under subtraction
8. All the integers under addition
9. All the integers under multiplication
10. All the positive real numbers under addition
11. All the positive real numbers under multiplication
12. All the positive real numbers under division
13. All the polynomials over \mathbb{Z}_2 under addition
14. All the polynomials over \mathbb{Z}_2 under multiplication

In each of the groups in Exercises 9.2.15 to 9.2.27 pair each element with its inverse.

15. $(\mathbb{Z}_4, +)$

19. $\mathbb{Z}_2[x, \leq 1]$

23. (\mathbb{Z}_6^*, \cdot)

16. (\mathbb{Z}_5^*, \cdot)

20. S_3

24. $\sqrt[6]{1}$

17. $\sqrt[4]{1}$

21. $(\mathbb{Z}_5, +)$

25. $(\mathbb{Z}_6, +)$

18. K

22. $\mathbb{Z}_2[x, \leq 2]$

26. $\mathbb{Z}_3[x, \leq 1]$

27. The Quaternion group.

28. Prove that if G is a group for which $(ab)^\# = a^\# b^\#$ for every pair of elements a, b , then G is a commutative group.

29. Prove that if G is a group in which every nonidentity element has order 2, then G is commutative.

30. What geometrical feature characterizes the multiplication table of commutative groups?

31. Explain why the Latin square below is not the multiplication table of a group.

1	2	3	4	5
2	1	5	3	4
3	4	1	5	2
4	5	2	1	3
5	3	4	2	1

32. Explain why the Latin square below is not the multiplication table of a group.

a	b	c	d	e	f
b	c	a	e	f	d
c	a	b	f	d	e
d	e	f	a	b	c
e	f	d	c	a	b
f	d	e	b	c	a

	Id	a	b	c
Id	Id	a	b	c
a	a	Id	c	b
b	b	c	Id	a
c	c	b	a	Id

Table 9.4 The Klein 4-group

	0	1	2	3
0	0	1	2	3
1	1	2	3	0
2	2	3	0	1
3	3	0	1	2

Table 9.5 Addition modulo 4

33. Prove that every 1-by-1 and every 2-by-2 Latin square is the multiplication table of a group.
34. Prove that every 3-by-3 Latin square is the multiplication table of a group.
35. Prove that every 4-by-4 Latin square is the multiplication table of a group.
36. Prove that if a is an element of the abstract group G , and m and n are arbitrary integers, then $a^m a^n = a^{m+n}$ and $(a^m)^n = a^{mn}$.
37. Prove that every group has a unique identity element.

9.3 Isomorphisms of Groups and Orders of Elements

Since abstract groups are defined in terms of their multiplication tables, it makes sense to identify abstract groups that have identical tables. With this in mind, let us examine Tables 9.4 to 9.7. Table 9.4 is an abbreviation of the table associated with the Klein 4-group. Table 9.5 represents $(\mathbb{Z}_4, +)$. Table 9.6 represents a mystery group, as yet unidentified. Table 9.7 represents the group $(F, +)$ where F is the Galois field $\text{GF}(2, x^2 + x + 1)$.

Table 9.4 and Table 9.7 are readily recognized as being essentially one and the same. In each table the diagonal entry is the group identity, and in each table the group multiplication of any two distinct nonidentity elements equals the third nonidentity element.

	e	A	B	C
e	e	A	B	C
A	A	e	C	B
B	B	C	A	e
C	C	B	e	A

Table 9.6 A mystery group

	0	1	α	$1 + \alpha$
0	0	1	α	$1 + \alpha$
1	1	0	$1 + \alpha$	α
α	α	$1 + \alpha$	0	1
$1 + \alpha$	$1 + \alpha$	α	1	0

Table 9.7 Addition in $\text{GF}(2, x^2 + x + 1)$

Table 9.5 and Table 9.6 are also essentially the same. To see this it is merely necessary to switch the columns and rows of Table 9.6 that correspond to the elements A and B so that this table takes the form displayed in Table 9.8.

When the symbols e , B , A , and C of Table 9.8 are replaced by 0, 1, 2, and 3, respectively, Table 9.5 is obtained, thus showing that Table 9.5 and Table 9.6 are only superficially different.

Is it possible that some such switching of rows and columns and rewriting of symbols could transform Table 9.4 to Table 9.5? The answer is no and a reason for this can be found in the diagonals of these tables. Notice that the diagonal of Table 9.4 contains only the group identity, whereas the diagonal of Table 9.5 contains another element besides the identity. Now it is clear that no matter how the symbols of Table 9.4 are relabeled, the diagonal will always contain only that symbol that stands for the group identity. Moreover, even when the row and column of any element are extracted and moved (in a consistent manner) to a new location, the diagonal still only contains the group identity. Hence, Table 9.4 and Table 9.5 are different in an essential way.

We formalize this notion of sameness with the term of isomorphism. Two groups (G, \cdot) and (H, \oplus) are said to be isomorphic provided that their elements can be matched up so that when the elements of G in a table of (G, \cdot) are replaced with the corresponding

	e	B	A	C
e	e	B	A	C
B	B	A	C	e
A	A	C	e	B
C	C	e	B	A

Table 9.8 A rewriting of the mystery group

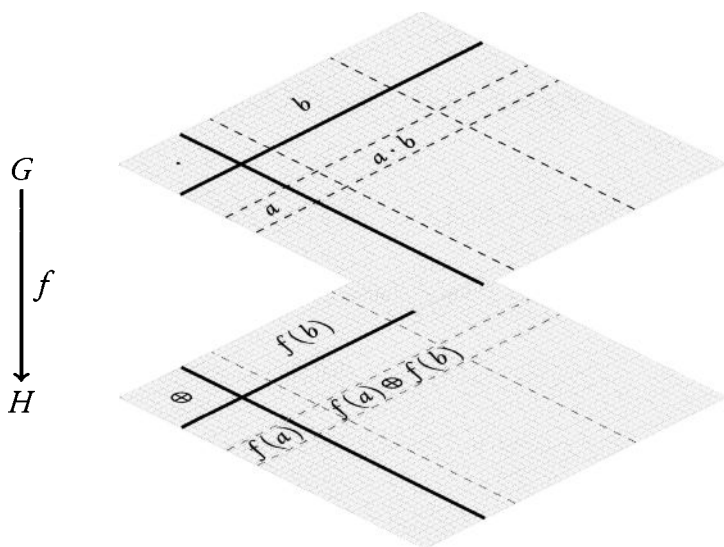


Figure 9.9 The third requirement for isomorphisms

elements of H , then the table of (G, \cdot) is transformed into a multiplication table for (H, \oplus) . In other words, the two groups (G, \cdot) and (H, \oplus) are isomorphic if there exists a function $f : G \rightarrow H$ such that

- (a) $f(a) = f(b)$ if and only if $a = b$,
- (b) for every $h \in H$ there is a $g \in G$ such that $f(g) = h$, and
- (c) $f(a \cdot b) = f(a) \oplus f(b)$.

The first requirement says that f assigns distinct elements of H to distinct elements of G . The second requirement says that every element of H is assigned to some element of G . When dealing with finite groups, which are our main concern here, these two conditions are redundant in the sense that each implies the other. For infinite groups this need not be the case (see Exercises 9.3.20 and 9.3.21).

To understand the last requirement, note that if in a multiplication table for (G, \cdot) each element a is replaced by $f(a)$, and if the result is a multiplication table for (H, \oplus) , then the entry $a \cdot b$ in the a -row and b -column will be replaced by $f(a \cdot b)$. However, because this is now the entry in the $f(a)$ -row and $f(b)$ -column of the multiplication table of (H, \oplus) , it must also equal $f(a) \oplus f(b)$ (see Figure 9.9). Hence, $f(a \cdot b) = f(a) \oplus f(b)$. The function f is called an *isomorphism*. Contrary to its usage in elementary calculus, the word function is used here in its most abstract sense, that of a mere association. Thus, the function

$$f(\text{Id}) = 0, \quad f(a) = 1, \quad f(b) = \alpha, \quad f(c) = 1 + \alpha$$

is an isomorphism of the group in Table 9.4 and the group in Table 9.7.

The function

$$f'(\text{Id}) = 0, \quad f'(a) = \alpha, \quad f'(b) = 1 + \alpha, \quad f'(c) = 1$$

is another isomorphism of the group in Table 9.4 and the group in Table 9.7. Similarly

$$g(e) = 0, \quad g(A) = 2, \quad g(B) = 1, \quad g(C) = 3$$

is an isomorphism of the group in Table 9.6 and the group in Table 9.5. However, the function

$$g'(e) = 0, \quad g'(A) = 1, \quad g'(B) = 2, \quad g'(C) = 3$$

is not an isomorphism, because it violates the last requirement of isomorphisms, since $g'(A \cdot A) = g'(e) = 0$ but $g'(A) + g'(A) = 1 + 1 \neq 0$.

If the groups G and H are *isomorphic groups* (are isomorphic to each other), then this fact is denoted by writing $G \cong H$. Some of the more general properties are given in Exercise 9.3.23.

It is clear that if two finite groups are isomorphic, then they must have the same number of elements. This leads us to define the *order of a finite group* as the number of its elements, so that $(\mathbb{Z}_n, +)$ has order n and S_n has order $n!$. In his note of 1878, reproduced in Appendix E, Cayley already mentions the problem of classifying all the isomorphism types of groups. The difficulty of this task is underscored by the fact that Cayley asserts erroneously that, up to isomorphism, there are three groups of order 6. It will be proved in Chapter 11 that every group of order 6 is isomorphic to either $(\mathbb{Z}_6, +)$ or to S_3 .

One useful tool for distinguishing between nonisomorphic groups is the property of commutativity. It is easily verified (Exercise 9.3.19) that for $n \geq 3$ the dihedral group D_n is not commutative, as is the case for the Quaternion group. Thus, the commutative group $(\mathbb{Z}_8, +)$ is not isomorphic to either D_4 or to the Quaternion group.

Another tool for distinguishing between nonisomorphic groups is provided by the notion of the order of an element of a group. In analogy with the notion of the order of a permutation, we define the *order* $o(a)$ of an element a of an abstract group G as the least positive integer n such that $a^n = 1_G$. If no such integer n exists, then we say that the element a has infinite order.

The element d of the Quaternion group has order 2, whereas each of the elements a, b, c, e, f , and g has order 4. The identity element always has order 1, and it is clearly the only group element that can have order 1. On the other hand, the element 2 of \mathbb{Z} has infinite order since $2 + 2 + 2 + \cdots + 2$ is never $0 = 1_{\mathbb{Z}}$.

Suppose now that the groups (G, \cdot) and (H, \oplus) are isomorphic. Since this is tantamount to saying that they have the same multiplication tables, it follows that corresponding elements of G and H must have the same orders.

Returning to the Quaternion group and the dihedral group D_4 , the first has exactly one element of order 2, whereas the latter has five such elements. Consequently, these two groups are not isomorphic.

Group-theoretic order enjoys the same properties as do the orders of roots of unity and permutations. Proposition 9.7 below is practically identical with Proposition 7.7, and for that reason no proof is offered.

Proposition 9.7 Let g and h be elements of a finite group G . Then,

- (a) $g^n = 1_G$ if and only if n is a multiple of $o(g)$;
- (b) $g^a = g^b$ if and only if $a - b$ is a multiple of $o(g)$;
- (c) if $o(g) = n$, then $o(g^k) = n/(k, n)$;
- (d) $o(gh) = o(g)o(h)$ if $o(g)$ and $o(h)$ are relatively prime and $gh = hg$.

That the assumption $gh = hg$ is necessary in the last part of Proposition 9.7 can be seen by choosing $g = (1\ 2)(3\ 4)$ and $h = (1\ 4\ 5)$. Then

$$o(gh) = o((1\ 2)(3\ 4)(1\ 4\ 5)) = o((1\ 3\ 4\ 5\ 2)) = 5 \neq 2 \cdot 3 = o(g)o(h).$$

This assumption is of course automatically satisfied in the context of fields wherein Proposition 7.7 was stated.

	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>
<i>a</i>	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>
<i>b</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>a</i>
<i>c</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>a</i>	<i>b</i>
<i>d</i>	<i>d</i>	<i>e</i>	<i>a</i>	<i>b</i>	<i>c</i>
<i>e</i>	<i>e</i>	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>

Table 9.9 A group table

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>
<i>A</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>
<i>B</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>A</i>
<i>C</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>A</i>	<i>B</i>
<i>D</i>	<i>D</i>	<i>E</i>	<i>A</i>	<i>B</i>	<i>C</i>
<i>E</i>	<i>E</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>

Table 9.10 A group table

Exercises 9.3

1. Prove that every two groups of order 2 are isomorphic to each other.
2. Prove that every two groups of order 3 are isomorphic to each other.
3. Are every two groups of order 4 isomorphic to each other?
4. Prove that the groups whose multiplication tables are in Table 9.9 and Table 9.10 are isomorphic to each other.
5. Prove that the groups whose multiplication tables are in Table 9.10 and Table 9.11 are isomorphic to each other.
6. Prove that the groups whose multiplication tables are in Table 9.9 and Table 9.11 are isomorphic to each other.
7. Explain why the cube and the octahedron of Figure 9.5 have isomorphic vertex groups.

	α	β	γ	δ	ε
α	α	β	γ	δ	ε
β	β	γ	δ	ε	α
γ	γ	δ	ε	α	β
δ	δ	ε	α	β	γ
ε	ε	α	β	γ	δ

Table 9.11 A group table

8. A double pyramid is formed by joining the vertices of a regular pentagon to two points, one directly above and one directly below the pentagon's geometrical center. Explain why the group of vertex symmetries of this solid is isomorphic to the dihedral group D_5 .

An isomorphism of a group with itself is called an *automorphism*.

9. Prove that the function $f(x) = 3x$ is an automorphism of $(\mathbb{Z}_{100}, +)$.
10. Prove that if k and n are relatively prime positive integers, then the function $f(x) = kx$ is an automorphism of $(\mathbb{Z}_n, +)$.
11. Is the function $f(x) = x^3$ an automorphism of $(\mathbb{Z}_{100}, +)$?
12. Is the function $f(x) = x^{-1}$ an automorphism of $(\mathbb{Z}_{17}^*, \cdot)$?
13. Prove that the function $f(x) = x^p$ is an automorphism of $(\text{GF}(p, P(x)), +)$.
14. Prove that the function $f(x) = x^p$ is an automorphism of $(\text{GF}(p, P(x))^*, \cdot)$.
15. For any element x of a group G let f_x be the function from G into itself defined by $f_x(a) = xax^{-1}$ for all $a \in G$. Prove that f_x is an automorphism of G .
16. For any element x of a group G let h_x be the function from G into itself defined by $h_x(a) = xax$ for all $a \in G$. Prove that h_x is an automorphism of G if and only if $x = x^{-1}$.
17. Prove that every group of even order contains an element of order 2.
18. Let G be a finite commutative group. The exponent of G is the least common multiple of all orders of the elements of G . Prove that G has an element whose order equals the exponent of G .
19. Prove that for $n \geq 3$ the dihedral group D_n is not commutative.

20. For the group $G = (\mathbb{Z}, +)$, find a function f from G to G that satisfies the first and third conditions but not the second condition.
21. Find a function f of the positive integers into themselves that satisfies the second condition for isomorphisms but does not satisfy the first condition.
22. Prove that if a and b are elements of a group G , then $o(a) = o(bab^{-1})$.
23. Let $f : G \rightarrow H$ be an isomorphism. Prove the following:
 - (a) $f(1_G) = 1_H$;
 - (b) for every $g \in G$, $f(g^{-1}) = [f(g)]^{-1}$.

9.4 Subgroups and Their Orders

A copy of the group

$$K = \{ \text{Id}, (1\ 2)(3\ 4), (1\ 3)(2\ 4), (1\ 4)(2\ 3) \},$$

consisting of the vertex symmetries of the rectangle, is clearly contained in the group

$$D_4 = \{ \text{Id}, (1\ 2)(3\ 4), (1\ 3)(2\ 4), (1\ 4)(2\ 3), (1\ 2\ 3\ 4), (1\ 4\ 3\ 2), (1\ 3), (2\ 4) \},$$

which consists of the vertex symmetries of the square. This relationship is formalized by the notion of a subgroup. If (G, \cdot) is an abstract group, and if H is a subset of G such that (H, \cdot) is a group in its own right, then (H, \cdot) is said to be a *subgroup* of (G, \cdot) . Thus, (K, \circ) is a subgroup of (D_4, \circ) . This is generally abbreviated to say that K is a subgroup of D_4 . Similarly, each of the groups $(\mathbb{Z}, +)$, $(\mathbb{Q}, +)$, $(\mathbb{R}, +)$, and $(\mathbb{C}, +)$ is a subgroup of the next. It is clear that if G is a group and H is a subset of G , then H is a subgroup of G if and only if the following conditions hold.

- (a) 1_G is in H ,
- (b) if a and b are in H , so is ab ,
- (c) if a is in H , then so is its inverse a^{-1} .

Thus $\{0, 2, 4\}$ is a subgroup of \mathbb{Z}_6 . If β is the Galois imaginary associated with the irreducible polynomial $x^3 + x^2 + 1$ over \mathbb{Z}_2 and $F = \text{GF}(2, x^3 + x^2 + 1)$, then each of the following sets defines a subgroup of $(F, +)$: $\{0, 1, \beta, 1 + \beta\}$, $\{0, 1, \beta^2, 1 + \beta^2\}$, and $\{0, \beta, \beta^2, \beta + \beta^2\}$.

The alternating group A_n that consists of all the even permutations of n symbols is a subgroup of the symmetric group S_n that consists of all the permutations on n symbols. The group of symmetries of the cube is a subgroup of the group S_8 of all the permutations

of the eight vertices of the cube. The group $\sqrt[6]{1} = \{1, -\omega^2, \omega, -1, \omega^2, -\omega\}$ contains both $\sqrt{1} = \{1, -1\}$ and $\sqrt[3]{1} = \{1, \omega, \omega^2\}$ as subgroups. If f is any function of n variables, then the group $S_{n,f}$ that consists of all the permutations that leave f unchanged is a subgroup of the symmetric group S_n .

Every group contains two obvious subgroups—itsself and the *trivial subgroup* $\{1_G\}$ that consists of the identity element of G alone. Any other subgroup of G is said to be *proper*.

The following theorem is quite possibly the most important theorem of group theory. In this form it was first stated and proved by Camille Jordan in his book *Traité des Substitutions*. Because he modestly attributed it to Lagrange who in fact had only proved the limited version of Corollary 9.13, this theorem nowadays bears the latter's name.

Theorem 9.8 (Lagrange's Theorem) If G is a finite group, then the order of any subgroup of G is a divisor of the order of G .

Since the proof of this theorem relies on the notion of a coset, a concept that is all but explicit in Lagrange's original proof, the proof is preceded by a discussion of this concept. If $H = \{h_1, h_2, h_3, \dots\}$ is any subgroup of the group (G, \cdot) , and if a is any element of the original group G , we define

$$a \cdot H = \{a \cdot h_1, a \cdot h_2, a \cdot h_3, \dots\}$$

and call $a \cdot H$ a *coset* of H . If $G = (\mathbb{Z}_{12}, +)$ and $H = \{0, 4, 8\}$, then

$$0 + H = \{0, 4, 8\} = H = 4 + H = 8 + H,$$

$$1 + H = \{1, 5, 9\} = 5 + H = 9 + H,$$

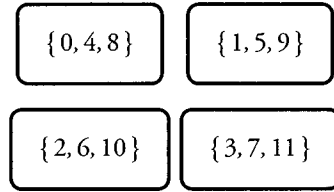
$$2 + H = \{2, 6, 10\} = 6 + H = 10 + H,$$

$$3 + H = \{3, 7, 11\} = 7 + H = 11 + H.$$

These cosets are pictured in Figure 9.10.

Another example begins with the polynomial $f = x_1 + x_2 - x_3 - x_4$ and its group

$$H = S_{4,f} = \text{Id}, (1\ 2), (3\ 4), (1\ 2)(3\ 4),$$

Figure 9.10 The cosets of $\{0, 4, 8\}$ in \mathbb{Z}_{12}

which is by definition a subgroup of S_4 . Here,

$$\begin{aligned}
 (1\ 2\ 3\ 4)H &= \{(1\ 2\ 3\ 4), (1\ 3\ 4), (1\ 2\ 3), (1\ 3)\} \\
 &= (1\ 3\ 4)H = (1\ 2\ 3)H = (1\ 3)H, \\
 (1\ 2\ 4\ 3)H &= (1\ 2\ 4\ 3), (1\ 4\ 3), (1\ 2\ 4), (1\ 4) = (1\ 4\ 3)H = (1\ 2\ 4)H = (1\ 4)H, \\
 (1\ 3\ 2\ 4)H &= \{(1\ 3\ 2\ 4), (1\ 4)(2\ 3), (1\ 3)(2\ 4), (1\ 4\ 2\ 3)\} \\
 &= (1\ 4)(2\ 3)H = (1\ 3)(2\ 4)H = (1\ 4\ 2\ 3)H, \\
 (1\ 3\ 4\ 2)H &= \{(1\ 3\ 4\ 2), (2\ 3\ 4), (1\ 3\ 2), (2\ 3)\} \\
 &= (2\ 3\ 4)H = (1\ 3\ 2)H = (2\ 3)H, \\
 (1\ 4\ 3\ 2)H &= \{(1\ 4\ 3\ 2), (2\ 4\ 3), (1\ 4\ 2), (2\ 4)\} \\
 &= (2\ 4\ 3)H = (1\ 4\ 2)H = (2\ 4)H.
 \end{aligned}$$

These sets are pictured in Figure 9.11.

The patterns that are indicated in the above examples hold in general.

Proposition 9.9 Let H be a subgroup of the group G . If H is finite, then every two cosets of H have the same number of elements, and every two distinct cosets of H are in fact disjoint.

Proof. It suffices to show that if H has m elements, then every coset of H also has m elements. Suppose $H = \{h_1, h_2, h_3, \dots, h_m\}$. Then $aH = ah_1, ah_2, ah_3, \dots, ah_m$. Moreover, if $ah_i = ah_j$, then $h_i = a^{-1}ah_i = a^{-1}ah_j = h_j$, and so distinct elements of H give rise to distinct elements of aH . Thus H and aH contain the same number of elements.

Suppose the two cosets aH and bH share some element. In other words, suppose there exist $h, k \in H$ such that $ah = bk$, or $a = bkh^{-1}$. Since H is a group in its own right, all the elements of the product $kh^{-1}H$ are back in H so that $kh^{-1}H \subset H$ and hence

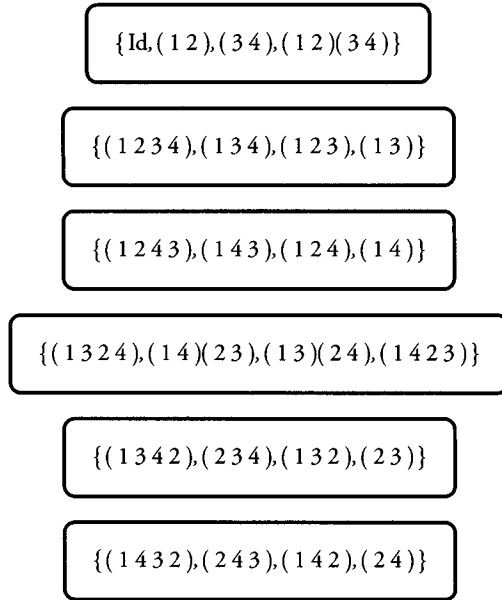


Figure 9.11 The cosets of $\{\text{Id}, (1\ 2), (3\ 4), (1\ 2)(3\ 4)\}$ in S_4

$aH = b(kh^{-1})H \subset bH$. A symmetrical argument leads to the inclusion $bH \subset aH$, and hence we may conclude that $aH = bH$. Thus, if any two cosets of H share an element, they must in fact be equal. In other words, distinct cosets must be disjoint. ■

Proof of Theorem 9.8. If H is subgroup of the finite group G , then, by Proposition 9.9, the cosets of H in G constitute a partition of the elements of G into sets all of which have the same cardinality as H . Consequently, the order of H must divide the order of G . ■

Section 7.2 contains such a computation of cosets. Specifically, it was demonstrated there that if γ is the Galois imaginary associated with $\text{GF}(2, x^4 + x^3 + x^2 + x + 1)$, then the cosets of the subgroup $\langle \gamma \rangle = \{1, \gamma, \gamma^2, \gamma^3, 1 + \gamma + \gamma^2 + \gamma^3\}$ of the multiplicative group $\text{GF}^*(2, x^4 + x^3 + x^2 + x + 1)$ consist of $\langle \gamma \rangle$ itself as well as the two sets

$$\{1 + \gamma, \gamma + \gamma^2, \gamma^2 + \gamma^3, 1 + \gamma + \gamma^2, \gamma + \gamma^2 + \gamma^3\}$$

and

$$\{1 + \gamma^2, \gamma + \gamma^3, 1 + \gamma + \gamma^3, 1 + \gamma^3, 1 + \gamma^2 + \gamma^3\}.$$

Cosets also played crucial, though implicit, roles elsewhere in this book. They appear in the proof of Galois's Theorem (Theorem 7.11). The sums used by Gauss in the proof of Theorem 2.15 are also cosets in disguise. Thus, the exponents of summands of A in this proof constitute the order 8 subgroup $\{1, 9, 13, 15, 16, 8, 4, 2\}$ of \mathbb{Z}_{17}^* and those of B are its only other coset. Also, the exponents of the summands of C constitute the subgroup $\{1, 13, 16, 4\}$ of \mathbb{Z} and those of D , E , and F are its cosets. Similarly, if z is any complex number, then the argument of z is in fact a coset of the subgroup

$$\langle 360^\circ \rangle = \{\dots, -720^\circ, -360^\circ, 0^\circ, 360^\circ, 720^\circ, \dots\}$$

of \mathbb{R} .

The number of cosets that the subgroup H has in G is called the *index* of H in G and is denoted by $[G : H]$. The corollary below follows directly from Proposition 9.9, seeing as $H = 1_G H$ is a coset.

Corollary 9.10 If H is a subgroup of the finite group G , then $[G : H]$ is equal to the order of G divided by the order of H .

The next corollary greatly facilitates the task of deciding when two elements belong to the same coset.

Corollary 9.11 Let H be a subgroup of G and let a and b be two elements of G . Then the following are equivalent:

- (a) $aH = bH$,
- (b) there exists an element c of G such that $a, b \in cH$,
- (c) $a^{-1}b \in H$.

Proof. (a) implies (b): Suppose $aH = bH$. Since $a = a1_G \in aH$ and $b = b1_G \in bH = aH$, it follows that both a and b belong to aH .

(b) implies (c): Suppose there is an element c of G such that $a, b \in cH$. In other words, suppose there exist $h, k \in H$ such that $a = ch$ and $b = ck$. Then,

$$a^{-1}b = (ch)^{-1}(ck) = h^{-1}c^{-1}ck = h^{-1}k \in H.$$

(c) implies (a): Suppose $a^{-1}b \in H$, or, in other words, $a^{-1}b = h$ for some $h \in H$. Then $b = ah$ and so the cosets aH and bH both contain the element $ah = b1_G$. By Proposition 9.9, $aH = bH$. ■

Not surprisingly, the significance of a coset depends on the meaning of both the ambient group and the defining subgroup. In the case of the subgroup

$$H = S_{4, x_1+x_2-x_3-x_4} = \{ \text{Id}, (1\ 2), (3\ 4), (1\ 2)(3\ 4) \}$$

of the symmetric group $G = S_4$, H consists of all the permutations of $\{1, 2, 3, 4\}$ that leave the polynomial $f = x_1 + x_2 - x_3 - x_4$ unchanged. Here, the cosets of H turn out to be in a one-to-one correspondence with the distinct variants of f . Thus, the elements of the coset

$$(1\ 2\ 3\ 4)H = \{ (1\ 2\ 3\ 4), (1\ 3\ 4), (1\ 2\ 3), (1\ 3) \}$$

all change f to the polynomial $x_2 + x_3 - x_4 - x_1$, the elements of the coset

$$(1\ 2\ 4\ 3)H = \{ (1\ 2\ 4\ 3), (1\ 4\ 3), (1\ 2\ 4), (1\ 4) \}$$

all change f to the polynomial $x_2 + x_4 - x_1 - x_3$, etc. In general we have the following proposition whose verification is relegated to Exercise 9.4.58.

Proposition 9.12 Let f be any function of the variables x_1, x_2, \dots, x_n . Then the two elements ρ and σ of S_n are in the same coset of $S_{n,f}$ if and only if $\rho f = \sigma f$.

We next point out what Lagrange actually proved.

Corollary 9.13 If f is a function of n variables and m is the number of distinct variants of f , then m is a divisor of $n!$.

Proof. According to Proposition 9.12, $m = [S_n : S_{n,f}]$, so that, by Corollary 9.10, m is a divisor of the order of S_n , which is $n!$. ■

Another setting wherein the cosets of a group have an interesting interpretation is that of the vertex symmetries of the tetrahedron. Let G be this group of vertex symmetries, and let H be the subgroup that consists of all the vertex symmetries that leave the vertex 4 unchanged. That is,

$$\begin{aligned} G = \{ & \text{Id}, (1\ 2)(3\ 4), (1\ 3)(2\ 4), (1\ 4)(2\ 3), (1\ 2\ 3), (1\ 3\ 2) \} \\ & \cup \{ (1\ 2\ 4), (1\ 4\ 2), (1\ 3\ 4), (1\ 4\ 3), (2\ 3\ 4), (2\ 4\ 3) \} \end{aligned}$$

and $H = \{\text{Id}, (1\ 2\ 3), (1\ 3\ 2)\}$. Then the cosets of H are

$$\begin{aligned} &\{\text{Id}, (1\ 2\ 3), (1\ 3\ 2)\}, && \{(1\ 2)(3\ 4), (2\ 4\ 3), (1\ 4\ 3)\}, \\ &\{(1\ 3)(2\ 4), (1\ 4\ 2), (2\ 3\ 4)\}, && \{(1\ 4)(2\ 3), (1\ 3\ 4), (1\ 2\ 4)\}. \end{aligned}$$

Note that the permutations of the first coset all fix 4, the permutations of the second coset all transform 4 to 3, those of the third coset all transform 4 to 2, and the permutations of the last coset all transform 4 to 1. There is a general principle in operation here, and it can be found in Exercise 9.4.40.

Exercises 9.4

1. Find all the subgroups of $(\mathbb{Z}_m, +)$ for $m = 1, 2, \dots, 10$.
2. Find all the subgroups of S_n for $n = 1, 2, 3$.
3. Show that S_4 contains subgroups of orders 1, 2, 3, 4, 6, 8, 12, and 24.
4. Show that S_4 contains two nonisomorphic subgroups of order 4.
5. Show that S_5 contains subgroups of orders 6, 8, 10, and 12.
6. Find all the subgroups of $(\text{GF}(2, x^2 + x + 1), +)$.
7. Find all the subgroups of $(\text{GF}(2, x^3 + x^2 + 1), +)$.
8. Find all the subgroups of D_n for $n = 3, 4, 5$.
9. Find all the subgroups of the Quaternion group.
10. Prove that A_4 does not have a subgroup of order 6.

Compute the cosets of the subgroup H of G as specified in Exercises 9.4.11 to 9.4.22.

11. $G = (\mathbb{Z}_{15}, +)$, $H = \{0, 3, 6, 9, 12\}$
12. $G = (\mathbb{Z}_{15}, +)$, $H = \{0, 5, 10\}$
13. $G = (\mathbb{Z}_{18}, +)$, $H = \{0, 9\}$
14. $G = (\mathbb{Z}_{24}, +)$, $H = \{0, 6, 12, 18\}$
15. $G = S_3$, $H = \{\text{Id}, (1\ 2)\}$
16. $G = S_3$, $H = \{\text{Id}, (1\ 2\ 3), (1\ 3\ 2)\}$
17. $G = A_4$, $H = K$
18. $G = A_4$, $H = \{\text{Id}, (123), (1\ 3\ 2)\}$
19. G is the Quaternion group, $H = \{1, d\}$
20. G is the Quaternion group, $H = \{1, a, d, e\}$
21. $G = D_n$, $H = \{1, \rho, \rho^2, \dots, \rho^{n-1}\}$, ρ is counterclockwise rotation by $2\pi/n$.
22. $G = D_{2n}$, $H = \{1, \alpha\}$, α is 180° rotation about any diagonal of the underlying polygon.

23. For which values of k can a function of three variables have k distinct variants? Justify your answer.
24. For which values of k can a function of four variables have k distinct variants? Justify your answer.
25. For which of the following values of k can a function of five variables have k distinct variants? Justify your answer.
- (a) $k = 1, 2, 3, 4, 5$ (b) $k = 10, 11, 12, 13, 14$
26. For which values of $k = 1, 2, 3, 4, 5, 6, 7$ can a function of 7 variables have k distinct variants? Justify your answer.
27. Prove that for each positive integer n there exists a function of n variables that has $(n-1)!$ variants.

In Exercises 9.4.28 to 9.4.37, does S_5 contain a subgroup of the given order? Justify your answers.

- | | | | | |
|--------|--------|--------|--------|-------|
| 28. 50 | 30. 30 | 32. 18 | 34. 10 | 36. 6 |
| 29. 40 | 31. 24 | 33. 15 | 35. 8 | 37. 5 |

38. Find a divisor d of $6! = 720$ such that S_6 does not have a subgroup of order d .
39. Prove that for every positive integer $n > 1$, the set A_n of even permutations of $\{1, 2, \dots, n\}$ is a subgroup of S_n . Show that A_n contains exactly half the elements of S_n .
40. Let G be a group of permutations of the set $\{1, 2, \dots, n\}$ and let G_1 be the set of all those permutations σ of G such that $\sigma(1) = 1$.
- (a) Prove that G_1 is a subgroup of G .
- (b) Show that two elements ρ and σ of G belong to the same coset of G_1 in G if and only if $\rho(1) = \sigma(1)$.
41. Suppose A and B are subgroups of G . Prove that $A \cap B$ is also a subgroup of G . Is the same true for $A \cup B$?
42. Suppose H is a subgroup of G and g is some element of G . Prove that the set $gHg^{-1} = \{ghg^{-1} \mid h \in H\}$ is also a subgroup of G and is isomorphic to H .

The *centralizer* Z_a of the element a of a group G consists of the set of all the elements of G that commute with a . That is, $Z_a = \{x \in G \mid xa = ax\}$.

43. Prove that for any a in G , Z_a is a subgroup of G .
 44. Compute the centralizer of each element of the following groups:

- | | |
|-------------------------|--------------------------|
| (a) $(\mathbb{Z}_n, +)$ | (c) A_4 |
| (b) S_3 | (d) the Quaternion group |

The *center* $Z(G)$ of the group G consists of the set of all the elements of G that commute with all the elements of G . That is, $Z(G) = \{x \in G \mid xa = ax \text{ for all } a \in G\}$.

45. Prove that the center of the group G is a subgroup of G .

Find the centers of the groups in Exercises 9.4.46 to 9.4.51.

- | | | |
|-------------------------|-----------|-----------|
| 46. $(\mathbb{Z}_n, +)$ | 48. S_3 | 50. D_4 |
| 47. Quaternion group | 49. A_4 | 51. S_n |

52. Let G be a permutation group on $\{1, 2, \dots, n\}$, and let H consist of all the even permutations of G . Prove that if H does not equal G , then its order equals half the order of G .
53. Prove that S_4 does not contain a subgroup isomorphic to the Quaternion group.
54. Prove that if $n > 4$ and p is a prime such that $p \leq n$, then S_n contains no subgroup of index k for any k such that $2 < k < p$.
55. Let $P(x)$ be a polynomial that is irreducible over \mathbb{Z}_p , where p is prime, and let $M(x)$ be the minimal polynomial of some $\alpha \in \text{GF}(p, P(x))$. Prove that the degree of $M(x)$ is a divisor of the degree of $P(x)$.
56. Let p be a prime number and n a positive integer. Prove that every group of order p^n has an element of order p .
57. Let f be a function of four variables that has three distinct variants. Prove that a certain pair of these variables can be interchanged without changing the value of the function. For example, if $f = xy + zw$, then $\{z, w\}$ constitute such a pair.
58. Prove Proposition 9.12.
59. Prove that A_5 does not contain a subgroup that is isomorphic to S_4 .
60. Prove that A_6 does not contain a subgroup that is isomorphic to S_5 .
61. Prove that the third condition for subgroups is redundant when G is finite.

62. Prove that if the union of two subgroups of G is a group, then one of those subgroups contains the other.

9.5 Cyclic Groups and Subgroups

If a is any element of the group G , we let $\langle a \rangle$ denote the set of all the integer powers of a . Since $a^0 = 1_G$, $a^m a^n = a^{m+n}$ for all integers m and n , and $(a^m)^\sharp = a^{-m}$, it follows that $\langle a \rangle$ is a subgroup of G . The subgroup $\langle a \rangle$ is said to be *generated* by a . If σ is a permutation, then there is nothing new about this notation; it was already used in the same sense in that more restricted context.

If $G = \mathbb{Z}$, then $\langle 2 \rangle$ consists of all the even integers. If $G = (\mathbb{Z}_{2n}, +)$, for some positive integer n , then $\langle 2 \rangle = \{0, 2, 4, \dots, 2n-2\}$ and $\langle n \rangle = \{0, n\}$. On the other hand, in \mathbb{Z}_5 , $\langle 2 \rangle = \{0, 2, 4, 1, 3\} = \mathbb{Z}_5$.

This, of course, generalizes to the fact that $\langle 2 \rangle = \mathbb{Z}_m$ whenever m is an odd integer. If ζ is the Galois imaginary associated with the irreducible polynomial $x^4 + x^3 + 1$ over \mathbb{Z}_2 , then ζ is primitive so that it has order 15 in F^* where $F = \text{GF}(2, x^4 + x + 1)$. Consequently, $\langle \zeta^3 \rangle = \{1, \zeta^3, \zeta^6, \zeta^9, \zeta^{12}\}$ and $\langle \zeta^5 \rangle = \{1, \zeta^5, \zeta^{10}\}$.

If it so happens that a is an element of the group G such that $\langle a \rangle = G$, then we say that a is a generator of G . Thus, 1 is always a generator of $(\mathbb{Z}_n, +)$, every odd element of \mathbb{Z}_8 is a generator of \mathbb{Z}_8 , and each of the elements $\zeta, \zeta^2, \zeta^4, \zeta^7, \zeta^8, \zeta^{11}, \zeta^{13}$, and ζ^{14} is a generator of the multiplicative group F^* of the Galois field above. A group that is generated by one of its elements is said to be a *cyclic group*. Thus, $(\mathbb{Z}_n, +)$ is cyclic for each n since, as noted above, it has 1 as a generator. The Primitive Element Theorem (Theorem 7.17) asserts that the multiplicative group of the nonzero elements of every Galois field is cyclic. On the other hand, the additive group of every Galois field $\text{GF}(p, P(x))$ is not cyclic for $p > 2$, since it has order p^v and every nonzero element has order p . Similarly, the group S_n is not cyclic for $n > 2$, since every cyclic group is necessarily commutative.

The cyclic groups are considered to be the simplest of all groups, and they can be classified in a very simple manner.

Theorem 9.14 Every two cyclic groups of the same order are isomorphic.

Proof. Let (G, \cdot) and (H, \oplus) be two finite cyclic groups of order n . Suppose that they are generated by the elements a and b , respectively, so that, because they are finite, $G = \{1_G, a, a^2, a^3, \dots, a^{n-1}\}$ and $H = \{1_H, b, b^2, b^3, \dots, b^{n-1}\}$. Then the function

$f(a^k) = b^k$ for each $k = 0, 1, 2, \dots, n-1$ is an isomorphism of G and H because

$$f(a^k \cdot a^m) = f(a^{k+m}) = b^{k+m} = b^k \oplus b^m = f(a^k) \oplus f(a^m)$$

where the exponents are added modulo n .

The proof of the theorem for infinite cyclic groups is relegated to Exercise 9.5.27. ■

It follows from this theorem that for a fixed positive integer n , the groups $(\sqrt[n]{1}, \cdot)$, $(\mathbb{Z}_n, +)$, $\langle (1 \ 2 \ \dots \ n) \rangle$ are all isomorphic to one another. Similarly, if $P(x)$ is irreducible of degree ν over \mathbb{Z}_p , then the group $(\text{GF}^*(p, P(x)), \cdot)$ is isomorphic to $(\mathbb{Z}_{p^\nu-1}, +)$. In particular, (\mathbb{Z}_p^*, \cdot) is isomorphic to $(\mathbb{Z}_{p-1}, +)$.

As was mentioned above, one of the main tasks of abstract group theory is the classification of all groups up to isomorphism. Since every two isomorphic groups necessarily have the same order, this can be rephrased as looking for the classification of all groups of a fixed order n . The next proposition resolves this classification problem when n is a prime number.

Proposition 9.15 Every group of a prime order p is cyclic and is therefore isomorphic to $(\mathbb{Z}_p, +)$.

Proof. Let G be a group of order p where p is a prime integer. Let a be any nonidentity element of G . If k is the order of a , then the subgroup $\langle a \rangle$ also has order k . However, by Theorem 9.8, k must divide p which is prime. Since $k > 1$, it follows that $k = p$, so that $\langle a \rangle = G$. ■

Clearly, then, every group of order 5 is necessarily isomorphic to $(\mathbb{Z}_5, +)$. Curiously, this fact seems to have eluded Cayley in his 1878 paper. Another surprising consequence of this proposition is that every group of prime order is necessarily commutative, since it is isomorphic to a commutative group. The following consequence is also very useful.

Proposition 9.16 If G is a group of finite order n , and if a is any element of G , then $a^n = 1_G$. Consequently, the order of a is a divisor of the order of G .

Proof. Let G be a group of order n , and let a be an element of G . The subgroup $\langle a \rangle$ has order $o(a)$. By Theorem 9.8, $o(a)$ is therefore a divisor of n . It now follows from Proposition 9.7 that $a^n = 1_G$. ■

Proposition 9.16 yields new proofs of Fermat's Theorem (Theorem 5.15) and Galois's Theorem (Theorem 7.11; see Exercise 9.5.12).

We conclude this section by pointing out that the information we have obtained so far also allows us to classify all the groups of order 4 up to isomorphism.

Proposition 9.17 If G is a group of order 4, then it is isomorphic to either $(\mathbb{Z}_4, +)$ or to K .

Sketch of proof. If G has an element of order 4, then it is cyclic and hence it is isomorphic to $(\mathbb{Z}_4, +)$. Otherwise, every nonidentity element of G has order 2, meaning that the diagonal entries of the multiplication table of G are all 1_G . It is now easily verified that the multiplication table of G must be identical with that of K . The details are relegated to Exercise 9.5.1. ■

With Proposition 9.17 available we have classified all the groups of order at most 5 up to isomorphism. Those of orders 1, 2, 3, and 5 are isomorphic to $(\mathbb{Z}_1, +)$, $(\mathbb{Z}_2, +)$, $(\mathbb{Z}_3, +)$, and $(\mathbb{Z}_5, +)$, respectively, whereas every group of order 4 is isomorphic to either $(\mathbb{Z}_4, +)$ or K . A sense of the enormity of the task of classifying all finite groups can be obtained from the amount of theory that was required by the classification of these small groups alone.

Exercises 9.5

1. Complete the proof of Proposition 9.17.

For each of the groups in Exercises 9.5.2 to 9.5.11, decide whether it is isomorphic to $(\mathbb{Z}_4, +)$ or to K .

2. $\{\text{Id}, (1\ 2), (3\ 4), (1\ 2)(3\ 4)\}$
3. (\mathbb{Z}_5^*, \cdot)
4. $\langle (1\ 2\ 3\ 4) \rangle$
5. $\langle (1\ 2\ 3\ 4)(5\ 6) \rangle$
6. $S_{4,f}$ where $f = x_1 + 2x^2 + 2x^3 + x^4$
7. (\mathbb{Z}_8, \cdot)
8. $(\text{GF}(2, x^2 + x + 1), +)$
9. $(\mathbb{Z}_{10}^*, \cdot)$
10. $\sqrt[4]{1}$
11. $\mathbb{Z}_2[x, \leq 1]$
12. Use Proposition 9.16 to give a new proof of Fermat's Theorem (Theorem 5.15) and Galois's Theorem (Theorem 7.11).
13. Prove that every subgroup of every cyclic group is also cyclic.
14. Suppose G is a group of order 187. Prove that if two subgroups of G have the same order, then they are isomorphic.

Find the largest value of k for which the groups in Exercises 9.5.15 to 9.5.23 contain a cyclic subgroup of order k .

- | | | |
|-------------------------------|--------------|----------------------------------|
| 15. $\sqrt[n]{1}$ | 18. S_5 | 21. A_{10} |
| 16. $(\text{GF}(p, P(x)), +)$ | 19. A_5 | 22. $(\mathbb{Z}_{37}^*, \cdot)$ |
| 17. D_{10} | 20. S_{10} | 23. $(\mathbb{Z}_{16}^*, \cdot)$ |
24. Prove that a group has exactly one subgroup if and only if it is isomorphic to $(\mathbb{Z}_1, +)$.
25. Prove that a group has exactly two subgroups if and only if it is isomorphic to $(\mathbb{Z}_p, +)$ for some prime p .
26. Prove that a group has exactly three subgroups if and only if it is isomorphic to $(\mathbb{Z}_{p^2}, +)$ for some prime p .
27. Complete the proof of Theorem 9.14 (the infinite case).
28. Prove that if m is relatively prime to n , then $m^{\varphi(n)} \equiv 1 \pmod{n}$ where $\varphi(n)$ is the Euler φ function.

9.6 Cayley's Theorem

The first groups to be examined by mathematicians were groups of permutations. It was not until a century had past that Cayley pointed out that every group is determined up to isomorphism by its multiplication table, and that therefore this table could be used to define the notion of an abstract group. At the same time Cayley noted that this innovation did not introduce any genuinely new structures into the study of groups, for, he said, every abstract group can be shown to be isomorphic to a group of permutations. Cayley did not formally prove this assertion; he contented himself with an example. His short note on the subject is included as Appendix E. Cayley's assertion will be formally stated and proved below as Theorem 9.18, but we first paraphrase Cayley's ideas in more modern terminology. Table 9.12 contains the multiplication table of the symmetric group S_3 with $a = (1\ 2)$, $b = (3\ 2\ 1)$, $c = (1\ 3)$, $d = (1\ 2\ 3)$, and $e = (2\ 3)$.

Reverting to our original notation for permutations, we associate with each element x a two-rowed array P_x whose first row is "Id $a\ b\ c\ d\ e$ " and whose second row is that row

	Id	a	b	c	d	e
Id	Id	a	b	c	d	e
a	a	Id	c	b	e	d
b	b	e	d	a	Id	c
c	c	d	e	Id	a	b
d	d	c	Id	e	b	a
e	e	b	a	d	c	Id

Table 9.12 The multiplication table of S_3

of Table 9.12 that corresponds to the element x . Thus,

$$\begin{aligned}
 P_{\text{Id}} &= \begin{pmatrix} \text{Id} & a & b & c & d & e \\ \text{Id} & a & b & c & d & e \end{pmatrix}, & P_a &= \begin{pmatrix} \text{Id} & a & b & c & d & e \\ a & \text{Id} & c & b & e & d \end{pmatrix}, \\
 P_b &= \begin{pmatrix} \text{Id} & a & b & c & d & e \\ b & c & d & a & \text{Id} & e \end{pmatrix}, & P_c &= \begin{pmatrix} \text{Id} & a & b & c & d & e \\ c & d & e & \text{Id} & a & b \end{pmatrix}, \\
 P_d &= \begin{pmatrix} \text{Id} & a & b & c & d & e \\ d & c & \text{Id} & e & b & a \end{pmatrix}, & P_e &= \begin{pmatrix} \text{Id} & a & b & c & d & e \\ e & b & a & d & c & \text{Id} \end{pmatrix}.
 \end{aligned}$$

Note that

$$P_a P_d = \begin{pmatrix} \text{Id} & a & b & c & d & e \\ a & \text{Id} & c & b & e & d \end{pmatrix} \begin{pmatrix} \text{Id} & a & b & c & d & e \\ d & c & \text{Id} & e & b & a \end{pmatrix} = \begin{pmatrix} \text{Id} & a & b & c & d & e \\ e & b & a & d & c & \text{Id} \end{pmatrix} = P_e = P_{ad}$$

and

$$P_d P_e = \begin{pmatrix} \text{Id} & a & b & c & d & e \\ d & c & \text{Id} & e & b & a \end{pmatrix} \begin{pmatrix} \text{Id} & a & b & c & d & e \\ e & b & a & d & c & \text{Id} \end{pmatrix} = \begin{pmatrix} \text{Id} & a & b & c & d & e \\ a & \text{Id} & c & b & e & d \end{pmatrix} = P_a = P_{de}.$$

In other words, the function that assigns to each element x the corresponding permutation P_x behaves just like an isomorphism. It is in fact always an isomorphism, and that is the gist of Cayley's assertion.

Theorem 9.18 (Cayley) Every group is isomorphic to a group of permutations.

Proof. Let G be any group. To each element x of G we assign a permutation P_x of the elements of G which transforms each element a to xa , i.e., $P_x(a) = xa$ for all $x, a \in G$. The function P_x is a permutation because $xa = xb$ if and only if $a = b$, and because

$P_x(x^{-1}y) = y$. Let H be the set of permutations thus obtained, i.e., $H = \{P_x \mid x \in G\}$. To see that H constitutes a group of permutations we observe first that for any two elements P_x and P_y of H we have

$$(P_x \circ P_y)(a) = P_x(P_y(a)) = P_x(ya) = xya = P_{xy}(a)$$

so that $P_x \circ P_y = P_{xy}$. Hence the composition of any two elements of H is in H , and the inverse of P_x is $P_{x^{-1}}$ which is also in H . Thus, H is a group of permutations.

Suppose x and y are distinct elements of G . Then

$$P_x(\text{Id}) = x\text{Id} = x \neq y = y\text{Id} = P_y(\text{Id})$$

and so P_x and P_y are distinct elements of H . Thus the function $f(x) = P_x$ matches up all the elements of G and H . That f is in fact an isomorphism follows from

$$f(xy) = P_{xy} = P_x \circ P_y = f(x) \circ f(y). \quad \blacksquare$$

Exercises 9.6

1. Write out the Cayley representation P_x for every element x of $(\mathbb{Z}_2, +)$.
2. Write out the Cayley representation P_x for every element x of $(\mathbb{Z}_3, +)$.
3. Write out the Cayley representation P_x for every element x of $(\mathbb{Z}_4, +)$.
4. Write out the Cayley representation P_x for every element x of K .
5. Write out the Cayley representation P_x for every element x of $(\mathbb{Z}_5, +)$.
6. Write out the Cayley representation P_x for every element x of the Quaternion group.
7. Write out the Cayley representation P_x for every element x of $(\mathbb{Z}_6, +)$.
8. Prove that for any finite group G and for any element x of G , all the cycles in the disjoint cycle representation of P_x have the same length. Explain why this common cycle length is necessarily a divisor of the order of G .
9. A set with a binary operation whose multiplication table is a Latin square is called a loop. The definition of the permutation P_x applies to loops as well as to groups. Prove that a loop X is a group if the set of permutations P_x with $x \in X$ is a group (under composition).

Chapter Summary

We have defined groups, both concretely, as permutations, and abstractly, in terms of axioms. Many of the algebraic systems examined in the earlier chapters are examples of such groups and this leads to a natural classification problem. The notion of isomorphism was defined to formalize the notion of “sameness” of groups. Some inroads were made on this difficult, if not impossible, classification problem by recognizing cyclic groups and using them to show that any two groups of the same prime order are necessarily isomorphic. Finally, we presented Cayley’s Theorem which asserts that every abstract group is in fact isomorphic to some group of permutations.

Chapter Review Exercises

Mark the following true or false.

1. The set of permutations that leave the function $(x_1 + x_2)(x_3 + x_4)$ unchanged is a group.
2. Every permutation belongs to some group.
3. There is no function f such that $S_{n,f} = D_n$.
4. No group of permutations is an abstract group.
5. $\sqrt[17]{1}$ is an abstract group.
6. $\sqrt[17]{1}$ is a permutation group.
7. The inverse of $(1\ 2\ 3\ 4\ 5)(6\ 7\ 8\ 9)$ is $(5\ 4\ 3\ 2\ 1)(9\ 8\ 7\ 6)$.
8. Every two groups of order 4 are isomorphic.
9. If the element a of the group G has order 24, then $\text{o}(a^{20}) = 6$.
10. S_6 has a subgroup of order 6.
11. S_6 has a subgroup of order 7.
12. If H is a subgroup of G , then every coset of H in G is also a subgroup of G .
13. If a subgroup of S_n has order n , then its index is $(n-1)!$.
14. The permutations $(1\ 2)$ and $(1\ 3)$ belong to the same coset of A_5 in S_5 .
15. Let $f = x_1x_2 + x_3 + x_4 + x_5$. Then the permutations $(1\ 3)(2\ 4\ 5)$ and $(1\ 4\ 5\ 2\ 3)$ belong to the same coset of $S_{5,f}$.
16. There is a function of 10 variables that has 11 distinct variants.
17. Every two groups of order 17 are isomorphic.

18. S_{10} has no element of order 11.
19. Given any three groups of order 4, some two of them are isomorphic.
20. The group $(\mathbb{Z}_{25}, +)$ is isomorphic to some group of permutations.

New Terms

abstract group, 193	index, 210
alternating group, 186	isomorphic groups, 202
automorphism, 205	isomorphism, 202
center, 214	Klein 4-group, 187
centralizer, 214	Latin square, 196
commutative group, 195	Quaternion group, 195
coset, 207	subgroup, 206
cyclic group, 215	symmetric group, 184
dihedral group, 188	trivial subgroup, 207
group of permutations, 184	vertex symmetries, 187

Supplementary Exercises

1. Write a computer script which will decide whether or not a given multiplication table is a Latin square.
2. Write a computer script which will decide whether or not a given multiplication table describes a group.
3. Write a computer script that will list all the subgroups of a group that is given by its multiplication table.
4. Write a computer script that will list all the subgroups of S_n for as many values of n as possible.
5. Classify the subgroups of S_n by isomorphism type for as many values of n as possible.

6. Decide whether the groups whose multiplication tables appear below are isomorphic.

<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>
<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>	<i>A</i>
<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>	<i>A</i>	<i>B</i>
<i>D</i>	<i>E</i>	<i>F</i>	<i>A</i>	<i>B</i>	<i>C</i>
<i>E</i>	<i>F</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
<i>F</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>

<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>
<i>b</i>	<i>c</i>	<i>a</i>	<i>e</i>	<i>f</i>	<i>d</i>
<i>c</i>	<i>a</i>	<i>b</i>	<i>f</i>	<i>d</i>	<i>e</i>
<i>d</i>	<i>e</i>	<i>f</i>	<i>a</i>	<i>b</i>	<i>c</i>
<i>e</i>	<i>f</i>	<i>d</i>	<i>b</i>	<i>c</i>	<i>a</i>
<i>f</i>	<i>d</i>	<i>e</i>	<i>c</i>	<i>a</i>	<i>b</i>

7. Characterize all the groups that have exactly four subgroups.
8. Characterize all the groups that have exactly five subgroups.
9. Write a computer program that will list all the cosets of a subgroup of a group.
10. For each positive integer n , what are the n -dimensional analogs of the five three-dimensional regular solids? Describe their vertex symmetry groups.

Chapter 10



QUOTIENT GROUPS AND THEIR USES

WE WILL DEFINE AN OPERATION ON GROUPS that is in some ways analogous to division of integers. This operation will provide us with a rigorous method for constructing the Galois fields of Chapter 7 as well as a host of new fields. Finally, Galois's theorem on the resolvability of algebraic equations will be described.

10.1 Quotient Groups

The notion of a set as a point is one of the recurrent themes of modern mathematics. Loosely speaking, the idea is to create new structures from old ones by considering sets of elements of the old structure as elements of the new structure. In the context of group theory this process takes the following form. Let (G, \cdot) be a group, and let S and T be any two subsets of G . We define

$$S \cdot T = \{ a \cdot b \mid a \in S \text{ and } b \in T \}$$

and $S^{-1} = \{ a^{-1} \mid a \in S \}$.

For example, if $G = (\mathbb{Z}_9, +)$, $S = \{1, 3, 5\}$, and $T = \{3, 7\}$, then

$$S + T = \{1 + 3, 1 + 7, 3 + 3, 3 + 7, 5 + 3, 5 + 7\} = \{4, 8, 6, 1, 8, 3\} = \{1, 3, 4, 6, 8\}$$

and

$$S^{-1} = \{-1, -3, -5\} = \{4, 6, 8\}.$$

Given a group (G, \cdot) , this process defines a binary operation on all the subsets of G . The associativity of the operation \cdot on the elements of G entails its associativity as an operation on sets as well (Exercise 10.1.35). Let us examine this operation on some

	H_0	H_1	H_2
H_0	H_0	H_1	H_2
H_1	H_1	H_2	H_0
H_2	H_2	H_0	H_1

Table 10.1 A quotient group of $(\mathbb{Z}_{12}, +)$

collections of cosets. Suppose $(G, \cdot) = (\mathbb{Z}_{12}, +)$, H is the subgroup $\{0, 3, 6, 9\}$, and let the cosets of H be labeled

$$H_0 = H = \{0, 3, 6, 9\},$$

$$H_1 = 1 + H = \{1, 4, 7, 10\},$$

$$H_2 = 2 + H = \{2, 5, 8, 11\}.$$

Then, for example,

$$\begin{aligned} H_1 + H_2 &= \{1+2, 1+5, 1+8, 1+11, 4+2, 4+5, 4+8, 4+11\} \\ &\cup \{7+2, 7+5, 7+8, 7+11, 10+2, 10+5, 10+8, 10+11\} \\ &= \{3, 6, 9, 0, 6, 9, 0, 3, 9, 0, 3, 6, 0, 3, 6, 9\} \\ &= \{0, 3, 6, 9\} = H_0. \end{aligned}$$

In fact, the sum of any two of the cosets of this subgroup H is again a coset of H . Table 10.1 summarizes the results of the operation of addition on the cosets of H and makes it clear that the cosets of H form a new group that is isomorphic to \mathbb{Z}_3 .

Let us examine the Quaternion group G of Table 9.3. Since $d^2 = 1$, it follows that $H = \{1, d\}$ is a subgroup of G . Its cosets are

$$1H = dH = \{1, d\},$$

$$aH = eH = \{a, e\},$$

$$bH = fH = \{b, f\},$$

$$cH = gH = \{c, g\}.$$

Note that

$$(aH)(bH) = \{ab, af, eb, ef\} = \{c, g, g, c\} = \{c, g\} = cH.$$

	H	aH	bH	cH
H	H	aH	bH	cH
aH	aH	H	cH	bH
bH	bH	cH	H	aH
cH	cH	bH	aH	H

Table 10.2 A quotient group of the Quaternion group

Table 10.2 displays the result of multiplying any two of the cosets of the subgroup $H = \{1, d\}$ of the Quaternion group. It is clear that this multiplication table (Table 10.2) is isomorphic with that of the Klein 4-group of Table 9.4. We hasten to point out that the cosets of a subgroup do not always form a group. For example, if $G = S_3$ and $H = \{\text{Id}, (1\ 2)\}$, then $(1\ 3)H = \{(1\ 3), (1\ 2\ 3)\}$, but

$$\begin{aligned} [(1\ 3)H][(1\ 3)H] &= \{(1\ 3)(1\ 3), (1\ 3)(1\ 2\ 3), (1\ 2\ 3)(1\ 3), (1\ 2\ 3)(1\ 2\ 3)\} \\ &= \{\text{Id}, (1\ 2), (2\ 3), (1\ 3\ 2)\} \end{aligned}$$

which is clearly not a coset of H . This counterexample notwithstanding, the previous two examples indicate that the cosets of a subgroup form a group of their own often enough for this phenomenon to merit attention. The following lemma acknowledges the fact that in one very special case, the product of two cosets is necessarily a coset.

Lemma 10.1 If H is any subgroup of G , then $HH = H$.

Proof. Since H is a subgroup, the product of any two of its elements is in H and hence, $HH \subseteq H$. On the other hand, since $1_G \in H$, it also follows that $HH \supset H1_G = H$. Hence, $HH = H$. ■

Let us examine the question of just when the product of two cosets is necessarily a coset. In order for the two cosets aH and bH of the subgroup H to have another coset, say cH , as their product, this coset cH would have to contain the element ab , since a is in aH and b is in bH . Hence, since the coset that contains ab is abH , we would have $(aH)(bH) = cH = abH$, or, upon multiplying both sides by $b^{-1}a^{-1}$, $b^{-1}HbH = H$. Hence for any $h \in H$,

$$b^{-1}hb = b^{-1}hb1_G \in b^{-1}HbH = H.$$

Consequently, if the product of every pair of cosets is again a coset, then $b^{-1}hb \in H$ for all $b \in G$ and $h \in H$.

This motivates the following definition. A subgroup H of G is said to be a *normal subgroup* if for every $b \in G$ and every $h \in H$, $b^{-1}hb \in H$. Thus, every subgroup of a commutative group G is normal since in such groups

$$b^{-1}hb = hb^{-1}b = h \in H.$$

If G is the Quaternion group, and if $H = \{1_G, d\}$, then clearly

$$x^{-1}1_Gx = x^{-1}x = 1_G$$

for every element $x \in G$. Moreover, since the column and the row of d in Table 9.3 are identical, it follows that for each element $x \in G$, $dx = xd$ and hence $x^{-1}dx = d$ for all such x . Thus $\{1, d\}$ is a normal subgroup of the Quaternion group.

It is clear that a subgroup H of G is normal if and only if $b^{-1}Hb \subset H$ for every element b of G .

Theorem 10.2 Let H be a subgroup of G . Then the following are equivalent:

- (a) H is a normal subgroup of G ;
- (b) $x^{-1}Hx = H$ for every element x of G ;
- (c) $xHx^{-1} = H$ for every element x of G ;
- (d) $Hx = xH$ for every element x of G ;
- (e) $(xH)(yH) = xyH$ for all $x, y \in G$;
- (f) the multiplication of the cosets of H forms a group.

Proof. (a) \implies (b). If H is a normal subgroup of G , then for every element x of G , $x^{-1}Hx \in H$. Consequently, if x is any element of G , two applications of the definition of normality yield

$$H \supset x^{-1}Hx \supset x^{-1}[(x^{-1})^{-1}Hx^{-1}]x = 1_GH1_G = H.$$

Consequently, $x^{-1}Hx = H$.

- (b) \implies (c). Replacing x by x^{-1} transforms the expression $x^{-1}Hx$ to xHx^{-1} .
- (c) \implies (d). If $xHx^{-1} = H$, then $xH = xH(x^{-1}x) = (xHx^{-1})x = Hx$.
- (d) \implies (e). Suppose $x, y \in G$. Since $Hx = xH$ we get

$$(xH)(yH) = x(Hy)H = x(yH)H = xyHH = xyH.$$

(e) \implies (f). Let H be a subgroup of G such that $(xH)(yH) = xyH$ holds for every two of its cosets. Then

$$(xH)(1_G H) = x1_G H = xH = 1_G xH = (1_G H)(xH)$$

so that the coset $H = 1_G H$ acts as the identity for this multiplication of cosets. In addition, for every coset xH we have

$$(xH)(x^{-1}H) = xx^{-1}H = 1_G H = H = x^{-1}xH = (x^{-1}H)(xH).$$

In other words, the coset $x^{-1}H$ is the inverse of the coset xH . As was noted above, the multiplication of subsets of a group is always associative. It follows that the multiplication of the cosets of H is indeed a group.

(f) \implies (a). This was already proved as part of the argument that motivated the definition of normal subgroups. ■

It would be useful to have a quick method for deciding whether a given subgroup is in fact normal. Unfortunately, no such method is known. There is, however, a variety of helpful ad hoc techniques. It is clear from the definition of normality that both the whole group G and the trivial subgroup $\{1_G\}$ are normal subgroups of G . More interestingly, we have the following observation.

Proposition 10.3 If H is a subgroup of index 2 in the group G , then H is a normal subgroup of G .

Proof. Assume for a contradiction that $xhx^{-1} \notin H$ for some $x \in G$ and $h \in H$. Then $x \notin H$, so that $xhx^{-1} \in xH$ and there exists an element k in H such that $xhx^{-1} = xk$. But then

$$x^{-1} = (xh)^{-1}(xk) = h^{-1}x^{-1}xk = h^{-1}k \in H,$$

contradicting the fact that $x \in H$. ■

The alternating group A_4 , which consists of all the even permutations in S_4 has twelve elements and it therefore contains exactly half the elements of the symmetric group S_4 . It follows that A_4 is a normal subgroup of S_4 which has only two cosets: itself and its complement in S_4 .

If H is a subgroup of G such that no other subgroup of G has the same order as H , then H is necessarily a normal subgroup of G (Exercise 10.1.21). Thus, since d is the

$H_0 = \{0, 3, 6, 9\}$	$H_1 = \{1, 4, 7, 10\}$	$H_2 = \{2, 5, 8, 11\}$
------------------------	-------------------------	-------------------------

Figure 10.1 The elements of $(\mathbb{Z}_{12}, +)/\{0, 3, 6, 9\}$

only element of the Quaternion group that has order 2, it follows that $\{1, d\}$ is the only subgroup of the Quaternion group that has order 2. It therefore is a normal subgroup of G .

If G is a commutative group, then for any subgroup H and for any element a of G , $a^{-1}Ha = Ha^{-1}a = H$, and so H is necessarily normal. This situation arises often enough to merit highlighting.

Proposition 10.4 Every subgroup of a commutative group is normal.

The converse to this proposition is false. The Quaternion group is an example of a noncommutative group all of whose subgroups are normal (Exercise 10.1.12).

In the special case where G is a permutation group, Exercise 8.2.23 provides an occasionally useful criterion for recognizing normal subgroups. According to this exercise, if ρ and σ are any two permutations in S_n , then $\rho\sigma\rho^{-1}$ and σ have the same number of k -cycles for each positive integer k . Consequently, the Klein 4-group

$$K = \{\text{Id}, (1\ 2)(3\ 4), (1\ 3)(2\ 4), (1\ 4)(2\ 3)\}$$

is a normal subgroup of A_4 , since K consists of the identity and all three of the elements of A_4 that consist of exactly two 2-cycles and nothing else.

The subgroup $H = \{\text{Id}, (1\ 2)\}$ of the symmetric group S_3 is not normal since for $x = (2\ 3)$ we have

$$xHx^{-1} = (2\ 3)\{\text{Id}, (1\ 2)\}(2\ 3) = \{\text{Id}, (1\ 3)\} \notin H.$$

If H is a normal subgroup of G , then the group formed by the cosets of H in G is called the *quotient* of G by H and is denoted by G/H . Each element of G/H is a coset of H in G and so it is a subset of G . Figures 10.1 and 10.2 illustrate this relationship for the examples displayed in Tables 10.1 and 10.2.

The nature of the quotient group G/H is also not easily determined, and again only ad hoc arguments are available. In some cases, however, the quotient group G/H is easily identified. If $H = G$, then H has only one coset in G , namely itself, and so G/H is



Figure 10.2 The elements of the Quaternion group/ $\{1, d\}$

necessarily the trivial group with one element only. At the other end of the spectrum we have the possibility that $H = \{1_G\}$. In this case, each coset of H consists of exactly one element of G , and so, in view of Theorem 10.2, G/H is isomorphic to G . A somewhat less trivial, though still surprisingly useful observation deals with subgroups of index 2, namely subgroups H that have exactly two cosets. As pointed out in Proposition 10.3, such subgroups are necessarily normal. Moreover, as they determine only two cosets, their quotients contain only two elements, and so they are isomorphic to $(\mathbb{Z}_2, +)$. Accordingly, S_4/A_4 is isomorphic to $(\mathbb{Z}_2, +)$.

Similarly, if H is a normal subgroup of index 3 in G , then G/H contains three elements and so it must be isomorphic to $(\mathbb{Z}_3, +)$. Since it was pointed out above that K is a normal subgroup of A_4 and since K has index 3 in A_4 , it follows that A_4/K is isomorphic to $(\mathbb{Z}_3, +)$.

We offer two observations that can be helpful in identifying quotient groups.

Proposition 10.5 Let H be a normal subgroup of the group G . Then the order of aH in G/H is a divisor of the order of a in G .

Proof. Suppose a has order m in G . Then

$$(aH)^m = (aH)(aH) \cdots (aH) = a^m H = 1_G H = H.$$

Since H is the identity element of the quotient group G/H , it follows from Proposition 9.7 that the order of aH in G/H is a divisor of m . ■

Consider the group $(F, +)$ where $F = \text{GF}(2, x^4 + x + 1)$. This group has order 16 and is commutative. If δ is the Galois imaginary that is associated with this field, then $H = \{0, 1, \delta, 1 + \delta\}$ is a subgroup of order 4. Since $(F, +)$ is commutative, this subgroup is normal. Thus, F/H is a group of order 4. Since every nonidentity element of F has order 2, it follows from the above proposition that every nonidentity element of F/H also has order 2. Hence, F/H is isomorphic to the Klein 4-group.

Proposition 10.6 If G is a cyclic group and H is a subgroup of G , then G/H is also cyclic.

Proof. Let g be a generator of G . Since the typical element of G has the form g^k , it follows that the typical element of G/H has the form $g^k H = (gH)^k$. Hence, G/H is generated by the coset gH , and so G/H is necessarily cyclic. ■

Consider the group (F^*, \cdot) where $F = \text{GF}(3, x^2 + x + 2)$ which was discussed in detail in Section 7.1. This is a commutative group of order 8. If δ is the Galois imaginary associated with this field, then we already know that δ is primitive so that δ is a generator of (F^*, \cdot) . Thus, (F^*, \cdot) is a cyclic group. This group has $H = \{1, \delta^4\}$ as a subgroup, which is necessarily normal. Since F^* was cyclic, so is F^*/H . Since the quotient F^*/H has order 4, it follows that F^*/H is isomorphic to $(\mathbb{Z}_4, +)$.

The quotient method for constructing groups provides us with a new perspective on modulo n arithmetic.

Corollary 10.7 Let $n\mathbb{Z}$ denote the subgroup of the group of integers \mathbb{Z} that is generated by the integer n . Then the quotient group $\mathbb{Z}/n\mathbb{Z}$ is isomorphic to $(\mathbb{Z}_n, +)$.

Proof. By definition,

$$n\mathbb{Z} = \{0, \pm n, \pm 2n, \pm 3n, \dots\},$$

and hence the cosets of H are $H, 1 + H, 2 + H, \dots, n - 1 + H$. Since, by Proposition 10.6 $\mathbb{Z}/n\mathbb{Z}$ is cyclic, it follows from Proposition 9.15 that $\mathbb{Z}/n\mathbb{Z}$ is isomorphic to $(\mathbb{Z}_n, +)$. ■

This corollary tells us that another way to define \mathbb{Z}_n is to view each of its elements as a coset of $n\mathbb{Z}$ in \mathbb{Z} . Thus, the element 2 in \mathbb{Z}_5 is to be regarded as a shorthand notation for the coset

$$2 + 5\mathbb{Z} = \{\dots, -8, -3, 2, 7, 12, \dots\}$$

of $5\mathbb{Z}$ in \mathbb{Z} .

However, the elements of \mathbb{Z}_n are also subject to multiplication, and it behooves us to verify that this multiplication is consistent with the coset point of view. To do this we define the product of the two cosets $a + n\mathbb{Z}$ and $b + n\mathbb{Z}$ in \mathbb{Z}_n as the coset $ab + n\mathbb{Z}$. There is a potential problem with this glib definition. Note that $2 + 5\mathbb{Z} = 7 + 5\mathbb{Z}$ and $4 + 5\mathbb{Z} = 9 + 5\mathbb{Z}$. Is the product of these two cosets to be $2 \cdot 4 + 5\mathbb{Z}$, $2 \cdot 9 + 5\mathbb{Z}$, $7 \cdot 4 + 5\mathbb{Z}$, or $7 \cdot 9 + 5\mathbb{Z}$? Fortunately, this question is moot since

$$2 \cdot 4 \equiv 2 \cdot 9 \equiv 7 \cdot 4 \equiv 7 \cdot 9 \pmod{5}$$

and so

$$2 \cdot 4 + 5\mathbb{Z} = 2 \cdot 9 + 5\mathbb{Z} = 7 \cdot 4 + 5\mathbb{Z} = 7 \cdot 9 + 5\mathbb{Z}.$$

This is the case in general as well. For, if both $a - a'$ and $b - b'$ are divisible by n , then

$$ab - a'b' = b(a - a') + a'(b - b')$$

is also divisible by n , and hence $ab + n\mathbb{Z} = a'b' + n\mathbb{Z}$.

Exercises 10.1

For each of the pairs G, H in Exercises 10.1.1 to 10.1.11, compute the cosets of H in G . Decide whether H is a normal subgroup of G . If it is normal, identify the isomorphism type of the quotient group G/H . If H is not a normal subgroup of G , explain why not.

- | | |
|---|--|
| 1. $G = (\mathbb{Z}_{16}, +)$, $H = \langle 4 \rangle$ | 7. $G = \sqrt[6]{1}$, $H = \{1, -1\}$ |
| 2. $G = (\mathbb{Z}_9, +)$, $H = \langle 3 \rangle$ | 8. $G = \sqrt[6]{1}$, $H = \{1, \omega, \omega^2\}$ |
| 3. $G = (\mathbb{Z}_{15}, +)$, $H = \langle 5 \rangle$ | 9. $G = A_4$, $H = \langle (1\ 2\ 3) \rangle$ |
| 4. $G = (\mathbb{Z}_{15}, +)$, $H = \langle 3 \rangle$ | 10. $G = (\mathbb{Z}_5^*, \cdot)$, $H = \langle 2 \rangle$ |
| 5. $G = S_3$, $H = \langle (1\ 2) \rangle$ | 11. $G = (\mathbb{Z}_{16}^*, \cdot)$, $H = \langle 7 \rangle$ |
| 6. $G = S_3$, $H = \langle (1\ 2\ 3) \rangle$ | |

For each of the groups G in Exercises 10.1.12 to 10.1.20 determine all the nontrivial normal subgroups H of G and identify G/H for such subgroup H .

- | | |
|--|--|
| 12. the Quaternion group | 17. $(\text{GF}^*(3, x^2 + x + 2), \cdot)$ |
| 13. D_4 | 18. $(\text{GF}(5, x^2 + 4x + 2), +)$ |
| 14. $(\text{GF}(2, x^2 + x + 1), +)$ | 19. A_4 |
| 15. $(\text{GF}(2, x^3 + x^2 + 1), +)$ | 20. D_5 |
| 16. $(\text{GF}(3, x^2 + x + 2), +)$ | |
21. Suppose that H is a subgroup of G such that no other subgroup of G is isomorphic to H . Prove that H is a normal subgroup of G .
 22. Suppose a is an element of the group G such that no other element has the same order as a . Prove that a has order 1 or 2, and that $\langle a \rangle$ is a normal subgroup of G .
 23. Prove that the center $Z(G)$ of the group G is a normal subgroup of G .

The element a of the group G is said to be *conjugate* to the element b if there exists an element x of G such that $xax^{-1} = b$. The set $C(a)$ consists of the set of all the elements of G that are conjugate to a and is called the *conjugacy class* of a .

24. Prove that a is conjugate to b if and only if b is conjugate to a .
25. Prove that if a is conjugate to b and b is conjugate to c , then a is conjugate to c .
26. Prove that if a and b are conjugate, then $C(a) = C(b)$.
27. Prove that if a and b are any two elements of a group G , then ab and ba are conjugate.
28. Describe the conjugacy class of each element of S_4 .
29. Describe the conjugacy class of each element of S_n .
30. Prove that the number of elements in $C(a)$ equals $[G : Z_a]$ and is therefore a divisor of the order of G whenever G is finite.
31. Prove that if p is a prime number, then every group of order p^n , $n > 0$, has a nontrivial center.
32. Prove that if H is a normal subgroup of G , then $(xH)^{-1} = x^{-1}H$ for all x in G .
33. Suppose H and K are normal subgroups of G such that $H \cap K = \{1_G\}$. Prove that $ab = ba$ whenever $a \in H$ and $b \in K$.
34. Prove that A_n is a normal subgroup of S_n and that $S_n/A_n \cong (\mathbb{Z}_2, +)$.
35. Let (G, \cdot) be any group. Prove that if A , B , and C are any subsets of G , then $A \cdot (B \cdot C) = (A \cdot B) \cdot C$.
36. Prove that the intersection of two normal subgroups of G is a normal subgroup of G .

10.2 Group Homomorphisms

Eventually it became convenient to express the properties of groups in terms of functions. Given two groups (G, \cdot) and (H, \oplus) and a function $f : G \rightarrow H$, f is said to be an isomorphism of these two groups (see Section 9.3) provided it satisfies the following two requirements:

- f is a bijection of G and H ;
- $f(a \cdot b) = f(a) \oplus f(b)$ for all $a, b \in G$.

A function $f : G \rightarrow H$ is said to be a *homomorphism* of G into H if it satisfies the second requirement (but not necessarily the first). We now offer a list of examples.

+	Even	Odd
Even	0	1
Odd	1	0

Table 10.3 A multiplication table

1. The function $f(a) = 2a$ is a homomorphism $(\mathbb{Z}, +) \rightarrow (\mathbb{Z}, +)$ because $f(a + b) = 2(a + b) = 2a + 2b = f(a) + f(b)$.

2. The function

$$f(a) = \begin{cases} 0 & \text{if } a \text{ is even;} \\ 1 & \text{if } a \text{ is odd} \end{cases}$$

is a homomorphism $(\mathbb{Z}, +) \rightarrow (\mathbb{Z}_2, +)$ as can be verified by means of Table 10.3 which summarizes the multiplication table of f .

3. The function $f(a) = 2a$ is a homomorphism $(\mathbb{Z}_2, +) \rightarrow (\mathbb{Z}_4, +)$. This is easily verified by the following observations:

$$f(0 + 0) = f(0) = 0 = 0 + 0 = f(0) + f(0);$$

$$f(0 + 1) = f(1) = 2 = 0 + 2 = f(0) + f(1);$$

$$f(1 + 0) = f(1) = 2 = 2 + 0 = f(1) + f(0);$$

$$f(1 + 1) = f(2) = 0 = 0 + 0 = f(1) + f(1).$$

In fact, if m is a positive integer, then the function $f_m(a) = ma$ is a homomorphism $(\mathbb{Z}, +) \rightarrow (\mathbb{Z}, +)$ because

$$f_m(a + b) = m(a + b) = ma + mb = f_m(a) + f_m(b).$$

A homomorphism $f : G \rightarrow H$ is said to be *injective* or a *monomorphism* provided the function $f : G \rightarrow H$ is injective. It is clear that the functions described in the first and third examples above are, in fact, monomorphisms, whereas the function of the second example is not.

4. If (A, \cdot) is a subgroup of (G, \cdot) , then the inclusion function defined by $i(a) = a \in A$ for all $a \in A$ is also a homomorphism since for all $a, b \in A$, $i(a + b) = a + b =$

$i(a) + i(b)$. This shows that the inclusion functions $\mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}$ are all (injective) homomorphisms.

5. If (G, \cdot) and (H, \oplus) are any two groups, then the function $f : G \rightarrow H$, defined by $f(a) = 1_H$ for all $a \in G$ is a homomorphism because

$$f(a \cdot b) = 1_b = 1_H \oplus 1_H = f(a) \oplus f(b).$$

6. Consider the function $f : \mathbb{Z}_4 \rightarrow \mathbb{Z}_4$ defined by $f(a) = 2a$. This is a homomorphism because for any two elements $a, b \in \mathbb{Z}_4$

$$f(a + b) \equiv 2(a + b) \equiv 2a + 2b \equiv f(a) + f(b) \pmod{4}.$$

Note that the codomain (or image) of this f is the proper subset $\{0, 2\}$ of \mathbb{Z}_4 .

7. If $(G, \cdot) = (S_n, \circ)$ and σ is any permutation in S_n , then the function defined as

$$f(\sigma) = \begin{cases} 0 & \text{if } \sigma \text{ is even;} \\ 1 & \text{if } \sigma \text{ is odd} \end{cases}$$

is a homomorphism. This follows from the definition of the parity of a permutation as the parity of the number of factors in any expression of the permutation as the composition of transpositions. The numerical value $f(\sigma)$ is called the *signature* of σ and is denoted by $\text{sign}(\sigma)$.

8. Let n be any fixed positive integer in \mathbb{Z} . Let \mathbb{T} denote the unit circle group under multiplication. The function $f_n : (\mathbb{Z}, +) \rightarrow \mathbb{T}$ defined as

$$f_n(k) = \cos \frac{2\pi k}{n} + i \sin \frac{2\pi k}{n}$$

is a homomorphism because

$$\begin{aligned} f_n(k + m) &= \cos \frac{2\pi(k + m)}{n} + i \sin \frac{2\pi(k + m)}{n} \\ &= \left(\cos \frac{2\pi k}{n} + i \sin \frac{2\pi k}{n} \right) \cdot \left(\cos \frac{2\pi m}{n} + i \sin \frac{2\pi m}{n} \right) = f_n(k) + f_n(m). \end{aligned}$$

The transition from the second equation above to the third one is not magic, it is De Moivre's Theorem.

Lemma 10.8 Let $f : (G, \cdot) \rightarrow (H, \oplus)$ be a homomorphism. Then $f(1_G) = 1_H$ and $f(a^{-1}) = (f(a))^{-1}$ for all $a \in G$.

Proof. Observe that

$$f(1_G) = f(1_G \cdot 1_G) = f(1_G) \oplus f(1_G) = (f(1_G))^2$$

and hence

$$1_H = f(1_G)(f(1_G))^{-1} = (f(1_G))^2(f(1_G))^{-1} = f(1_G),$$

proving the first equation.

Let a be any element of G ; then

$$1_H = f(1_G) = f(a \cdot a^{-1}) = f(a) \oplus f(a^{-1})$$

so that $f(a^{-1}) = (f(a))^{-1}$. ■

Let $f : G \rightarrow H$ be a group homomorphism. The *kernel* of f is denoted by $\text{Ker } f$ and is defined as

$$\text{Ker } f = f^{-1}(1_H) = \{ g \in G \mid f(g) = 1_H \}.$$

It is clear that in examples 1, 3, 4, and 8 above the kernel of f is a trivial group consisting of 1_G alone. However, in example 6 the kernel equals $\{0, 2\}$. In example 4 the kernel of f is the set $\{0, 2\}$. The kernel for example 7 is the alternating group A_n . In Example 6 the kernel of f_n is the infinite set

$$\{ \cos 2\pi mn + i \sin 2\pi mn \mid m \in \mathbb{Z} \}.$$

The examples above indicate that $\text{Ker } f$ is a subgroup of G . In fact more is true.

Proposition 10.9 (First Isomorphism Theorem) Let $f : G \rightarrow H$ be a group homomorphism. Then

- (a) $\text{Ker } f$ is a normal subgroup of G ;
- (b) $\text{Im } f$ is a subgroup of H ;
- (c) $\text{Im } f$ is isomorphic to the quotient group $G/\text{Ker } f$.

Proof. For (a), let a and b be elements of $\text{Ker } f$. Then

$$f(a \cdot b) = f(a) \oplus f(b) = 1_H \oplus 1_H = 1_H$$

implying that $a \cdot b \in \text{Ker } f$, which proves that the kernel is closed with respect to the group operations in G .

Now let c be any element of $\text{Ker } f$. Then c^{-1} exists in G . Hence

$$1_H = f(c) \oplus (f(c))^{-1} = 1_H \oplus f(c^{-1}) = f(c^{-1})$$

and hence $c^{-1} \in \text{Ker } f$.

Finally we address the normality issue. By Theorem 10.2 it suffices to show that for every element c of G ,

$$c^{-1} \cdot \text{Ker } f \cdot c \subset \text{Ker } f.$$

For this it suffices to show that for any such c and any $x \in \text{Ker } f$

$$c^{-1} \cdot x \cdot c \in \text{Ker } f.$$

This, however, is clear since

$$f(c^{-1} \cdot x \cdot c) = f(c^{-1}) \oplus f(x) \oplus f(c) = (f(c))^{-1} \oplus 1_H \oplus f(c) = (f(c))^{-1} \oplus f(c) = 1_H.$$

For (b), see Exercise 10.2.19.

For (c), let $K = \text{Ker } f$. Since $K = \text{Ker } f$ is now known to be a normal subgroup of G , the cosets of K in G can be written down as $\{aK = Ka \mid a \in G\}$. Keep in mind that $aK = bK$ if and only if $a^{-1}b \in K$. Let $\varphi: G/K \rightarrow H$ be defined via $\varphi(aK) = f(a)$. This function φ is well defined, for, if a and b are elements of the same cosets of K , then there exists an element k such that $b = ak$. Consequently

$$\varphi(bK) = f(b) = f(ak) = \varphi(akK) = \varphi(aK).$$

Moreover, for any two cosets aK and bK

$$\varphi(aK bK) = \varphi(abK) = f(ab) = f(a)f(b) = \varphi(aK)\varphi(bK)$$

so that $\varphi: G/K \rightarrow H$ is a homomorphism.

If h is an arbitrary element of H , then, because f is surjective, it follows that there exists an element $a \in G$ such that $f(g) = h$. Then $\varphi(gK) = f(g) = h$, which implies that φ is surjective.

Finally, for any elements a, b in G $\varphi(aK) = \varphi(bK)$ implies that $f(a) = f(b)$, or $[f(b)]^{-1}f(a) = 1_H$, or $f(b^{-1}a) = f(b^{-1})f(a) = 1_H$, so $b^{-1}a \in K$, which means that a and b are in the same coset of K . Thus φ is injective. ■

The next proposition can be viewed as a converse of the First Isomorphism Theorem.

Proposition 10.10 Let (A, \cdot) be a normal subgroup of (G, \cdot) . Then there exists a surjective homomorphism $f : (G, \cdot) \rightarrow (H, \oplus)$ such that $\text{Ker } f = A$.

Proof. This follows immediately from Theorem 10.2. ■

For example, the function associated with the quotient group displayed in Figure 10.1 is

$$\begin{aligned} f(0) &= f(3) = f(6) = f(9) = H_0, \\ f(1) &= f(4) = f(7) = f(10) = H_1, \\ f(2) &= f(5) = f(8) = f(11) = H_2. \end{aligned}$$

The function associated with the quotient group displayed in Figure 10.2 is

$$\begin{aligned} f(1) &= f(d) = H, & f(a) &= f(e) = aH, \\ f(b) &= f(f) = bH, & f(c) &= f(g) = cH. \end{aligned}$$

Suppose we wish to find all the homomorphisms $f : K \rightarrow S_3$ where K is the Klein 4-group. The subgroups of K are K , $\{1, a\}$, $\{1, b\}$, $\{1, c\}$, and $\{1\}$, and all of them are normal in K because K is abelian. Consequently, there are three homomorphisms that satisfy the requirements, namely $K/K \cong \{1\}$, $K/\{1, a\} \cong K/\{1, b\} \cong K/\{1, c\}$, and $K/\{1\} \cong K$. For $f : S_3 \rightarrow K$, the subgroups of S_3 are S_3 , A_3 , $\{1, a\}$, $\{1, b\}$, $\{1, c\}$, and $\{1\}$, but of these only S_3 , A_3 , and $\{1\}$ are normal. Hence there are three homomorphisms that satisfy the requirements.

Exercises 10.2

In each of Exercises 10.2.1 to 10.2.18, find all of the homomorphisms $f : G \rightarrow H$ where G and H are the given groups.

1. G is the trivial group and H is an arbitrary group.
2. G is an arbitrary group and H is the trivial group.
3. $G = H = (\mathbb{Z}_2, +)$ 11. $G = (\mathbb{Z}_5, +), H = (\mathbb{Z}_5, +)$
4. $G = (\mathbb{Z}_2, +), H = (\mathbb{Z}_3, +)$ 12. $G = (\mathbb{Z}_5, +), H = (\mathbb{Z}_3, +)$
5. $G = (\mathbb{Z}_3, +), H = (\mathbb{Z}_2, +)$ 13. $G = (\mathbb{Z}_2, +), H = (\mathbb{Z}_6, +)$
6. $G = (\mathbb{Z}_2, +), H = (\mathbb{Z}_4, +)$ 14. $G = (\mathbb{Z}_3, +), H = (\mathbb{Z}_6, +)$
7. $G = (\mathbb{Z}_3, +), H = (\mathbb{Z}_3, +)$ 15. $G = (\mathbb{Z}_6, +), H = (\mathbb{Z}_2, +)$
8. $G = (\mathbb{Z}_4, +), H = (\mathbb{Z}_2, +)$ 16. $G = (\mathbb{Z}_6, +), H = (\mathbb{Z}_3, +)$
9. $G = (\mathbb{Z}_2, +), H = (\mathbb{Z}_5, +)$ 17. $G = (\mathbb{Z}_4, +), H = \text{Klein 4-group}$
10. $G = (\mathbb{Z}_3, +), H = (\mathbb{Z}_5, +)$ 18. $G = \text{Klein 4-group}, H = (\mathbb{Z}_4, +)$
19. Prove the second part of the First Isomorphism Theorem.

10.3 The Rigorous Construction of Fields

If H is a normal subgroup of G , then, loosely speaking, we say that G/H inherits its binary operation from that of G . Some well-known groups, however, are subject to additional operations, as is the case, for instance, for the groups $(\mathbb{Z}, +)$ and $(F[x], +)$ for any field F . We saw above that the multiplication of integers could be transferred to $(\mathbb{Z}_n, +)$ so as to convert it to arithmetic modulo n . It will now be demonstrated that a similar procedure can be employed to yield rigorous constructions of both the complex numbers of Chapter 2 and the Galois fields of Chapter 7, as well as a host of new fields.

Let F be an arbitrary field and let $P(x)$ be any polynomial in $F[x]$. If we set $G = (F[x], +)$ and let H be the set of all the polynomials in $F[x]$ that are divisible by $P(x)$, then H is a subgroup of G . The reason for this is that $0 = 0 \cdot P(x)$ is clearly a multiple of $P(x)$, and if $A(x)P(x)$ and $B(x)P(x)$ are any two multiples of $P(x)$, then so are

$$A(x)P(x) + B(x)P(x) = (A(x) + B(x))P(x)$$

and $-A(x)P(x) = (-A(x))P(x)$ multiples of $P(x)$.

Since G is a commutative group it follows that H is necessarily a normal subgroup of G and so there exists a quotient group G/H . We shall show that when the polynomial $P(x)$ is irreducible the quotient group G/H can be converted to a field. The fields so obtained include the Galois fields as well as a variety of new fields. Consider the case where $F = \mathbb{Z}_2$ and $P(x) = x^2 + x + 1$ which is known to be irreducible over F . By Corollary 9.11, the polynomials $M(x)$ and $N(x)$ in $F[x]$ belong to the same coset of H if and only if

$$-N(x) + M(x) = M(x) - N(x) \in H$$

which is true if and only if $P(x)$ divides $M(x) - N(x)$. Since the difference of any two of the polynomials $0, 1, x, 1 + x$ has degree at most 1, it follows that the four cosets

$$0 + H, 1 + H, x + H, 1 + x + H \quad (10.11)$$

are all distinct. Moreover, if $M(x)$ is any polynomial of $F[x]$, then, by Proposition 6.4, there exist polynomials $Q(x)$ and $R(x)$ such that

$$M(x) = Q(x)(x^2 + x + 1) + R(x) \quad (10.12)$$

and $R(x)$ is either 0 or else has degree at most 1. In other words, $R(x)$ is one of the polynomials $0, 1, x$, or $1 + x$.

However, it is clear from Equation 10.12 above that $M(x) - R(x)$ is divisible by $x^2 + x + 1$ and hence $M(x)$ and $R(x)$ belong to the same coset of H . Consequently, every polynomial of $F[x]$ belongs to one of the cosets in List 10.11. Thus, G/H has four elements only. (It happens to be isomorphic to the Klein 4-group, but that is of no interest to us now.)

We shall next convert G/H , which already has the additive operation,

$$(a + H) + (b + H) = (a + b) + H$$

to a field by defining a multiplication of its elements. Specifically, we define

$$(a + H) \cdot (b + H) = ab + H.$$

These two operations on G/H can be tabulated; see Tables 10.4 and 10.5. In creating these tables we wrote the entry for $(x + H) \cdot (x + H) = x^2 + H$ as $1 + x + H$ since these

+	H	$1 + H$	$x + H$	$1 + x + H$
H	H	$1 + H$	$x + H$	$1 + x + H$
$1 + H$	$1 + H$	H	$1 + x + H$	$x + H$
$x + H$	$x + H$	$1 + x + H$	H	$1 + H$
$1 + x + H$	$1 + x + H$	$x + H$	$1 + H$	H

Table 10.4 Addition in $\mathbb{Z}_2[x]/(x^2 + x + 1)$

\cdot	H	$1 + H$	$x + H$	$1 + x + H$
H	H	H	H	H
$1 + H$	H	$1 + H$	$x + H$	$1 + x + H$
$x + H$	H	$x + H$	$1 + x + H$	$1 + H$
$1 + x + H$	H	$1 + x + H$	$1 + H$	$x + H$

Table 10.5 Multiplication in $\mathbb{Z}_2[x]/(x^2 + x + 1)$

two cosets are identical. These tables are in fact identical with those of $\text{GF}(2, x^2 + x + 1)$ (Tables 10.6 and 10.7) if the H is either suppressed or replaced by zero and the x is replaced by α .

Thus, our quotient G/H is none other than the Galois field $\text{GF}(2, x^2 + x + 1)$ in disguise. This laborious reconstruction of this Galois field has the advantage of mathematical rigor. Galois's construction of his fields presumed the existence of at least one zero for every irreducible polynomial. The only justification given by Galois for this assumption was the analogy with the complex numbers. The construction given here, on the other hand, makes no such assumptions.

The same method that was used above to construct $\text{GF}(2, x^2 + x + 1)$ can be applied in other situations to construct a wide variety of both old and new fields. Before describing this construction in its full generality, it is necessary to formalize some concepts. Two fields F and F' are said to be isomorphic if there is a function $f : F \rightarrow F'$ such that

- $f(a_1) \neq f(a_2)$ for any distinct $a_1, a_2 \in F$,
- for each $b \in F'$ there is an $a \in F$ such that $f(a) = b$,
- $f(a_1 + a_2) = f(a_1) + f(a_2)$ for any $a_1, a_2 \in F$,
- $f(a_1 a_2) = f(a_1) f(a_2)$ for any $a_1, a_2 \in F$.

+	0	1	α	$1+\alpha$
0	0	1	α	$1+\alpha$
1	1	0	$1+\alpha$	α
α	α	$1+\alpha$	0	1
$1+\alpha$	$1+\alpha$	α	1	0

Table 10.6 Addition in $\text{GF}(2, x^2 + x + 1)$

\cdot	0	1	α	$1+\alpha$
0	0	0	0	0
1	0	1	α	$1+\alpha$
α	0	α	$1+\alpha$	1
$1+\alpha$	0	$1+\alpha$	1	α

Table 10.7 Multiplication in $\text{GF}(2, x^2 + x + 1)$

The function f is said to be an isomorphism. A comparison of this notion with that of the group isomorphism defined in Section 9.3 leads to the conclusion that every field isomorphism of F and F' constitutes both a group isomorphism of $(F, +)$ and $(F', +)$ on the one hand and a group isomorphism of (F^*, \cdot) and (F'^*, \cdot) on the other. The function

$$f(r + s\alpha) = (r + sx) + H$$

constitutes an isomorphism of $\text{GF}(2, x^2 + x + 1)$ and the above constructed quotient G/H . The first two properties are clearly satisfied whereas the last two properties follow from an examination of the above tables and the following arguments:

$$\begin{aligned}
 f(r_1 + s_1\alpha + r_2 + s_2\alpha) &= f[(r_1 + r_2) + (s_1 + s_2)\alpha] \\
 &= (r_1 + r_2) + (s_1 + s_2)x + H \\
 &= (r_1 + s_1x + H) + (r_2 + s_2x + H) \\
 &= f(r_1 + s_1\alpha) + f(r_2 + s_2\alpha)
 \end{aligned}$$

and, since $\alpha^2 = \alpha + 1$ and $x^2 + H = x + 1 + H$,

$$\begin{aligned}
 f((r_1 + s_1\alpha)(r_2 + s_2\alpha)) &= f[r_1 r_2 + (r_1 s_2 + r_2 s_1)\alpha + s_1 s_2 \alpha^2] \\
 &= f[(r_1 r_2 + s_1 s_2) + (r_1 s_2 + r_2 s_1 + s_1 s_2)\alpha] \\
 &= (r_1 r_2 + s_1 s_2) + (r_1 s_2 + r_2 s_1 + s_1 s_2)x + H \\
 &= r_1 r_2 + (r_1 s_2 + r_2 s_1)x + s_1 s_2 x^2 + H \\
 &= (r_1 + s_1 x)(r_2 + s_2 x) + H \\
 &= (r_1 + s_1 x + H) \cdot (r_2 + s_2 x + H) \\
 &= f(r_1 + s_1 \alpha) \cdot f(r_2 + s_2 \alpha).
 \end{aligned}$$

We now generalize this procedure to arbitrary fields and irreducible polynomials, first addressing the issue of addition, and only later dealing with multiplication.

If $P(x)$ is any polynomial over the field F , then we denote by $(P(x))$ the set of all the polynomials of $F[x]$ that are divisible by $P(x)$. The set $(P(x))$ is a normal subgroup of $(F[x], +)$ and hence it determines a quotient group $F[x]/(P(x))$. The cosets of this group all have the unwieldy form $M(x) + (P(x))$, and it is convenient to replace this clumsy expression by $[M(x)] = M(x) + (P(x))$. It may be easily verified that in terms of this notation the addition of cosets assumes the form

$$[M(x)] + [N(x)] = [M(x) + N(x)]. \quad (10.13)$$

The following proposition makes it easy to visualize the additive group $F[x]/(P(x))$.

Proposition 10.14 Let F be a field and let $P(x)$ be a polynomial of degree d over F . Then $F[x]/(P(x))$ is the set of $[R(x)]$ such that $R(x) \in F[x]$ and either the degree of $R(x)$ is less than d or $R(x) = 0$.

Proof. By definition, every element of $F[x]/(P(x))$ has the form $[M(x)]$ for some polynomial $M(x)$ over F . If $M(x)$ is now divided by $P(x)$ so as to yield

$$M(x) = Q(x)P(x) + R(x)$$

with $R(x)$ being either the zero polynomial or having degree less than d , then, since

$$M(x) - R(x) = Q(x)P(x),$$

which is divisible by $P(x)$, it follows that $[M(x)] = [R(x)]$. In other words, every element of the quotient structure $F[x]/(P(x))$ can be represented in the form $[R(x)]$ where $R(x)$ is either the zero polynomial or else has degree less than d .

Suppose now that $R(x)$ and $R'(x)$ are two polynomials of degree less than d such that

$$[R(x)] = [R'(x)].$$

It then follows that the difference $R(x) - R'(x)$ is a polynomial of degree less than d that is divisible by the polynomial $P(x)$ of degree d . Clearly then, $R(x) - R'(x)$ must be the zero polynomial; in other words, $R(x) = R'(x)$.

Hence we have shown that the nonzero elements of $F[x]/(P(x))$ are in a one to one correspondence with the polynomials of degree less than d over F . ■

Accordingly,

$$\mathbb{Z}_2[x]/(x^2 + x + 1) = \{ [0], [1], [x], [1 + x] \},$$

$$\mathbb{Z}_2[x]/(x^2 + 1) = \{ [0], [1], [x], [1 + x] \},$$

$$\mathbb{Z}_3[x]/(x^2 + 1) = \{ [0], [1], [2], [x], [1 + x], [2 + x], [2x], [1 + 2x], [2 + 2x] \},$$

$$R[x]/(x^2 + 1) = \{ [a + bx] \mid a, b \in R \}.$$

Some puzzlement may be caused by the similarity between $\mathbb{Z}_2[x]/(x^2 + x + 1)$ and $\mathbb{Z}_2[x]/(x^2 + 1)$, since both have Table 10.4 as their addition tables. In fact, as additive groups these two structures are indeed isomorphic. It is only when multiplication is also brought into play that the difference between them becomes evident. The additive group $F[x]/(P(x))$ is endowed with an operation of multiplication by the following definition:

$$(Q(x) + (P(x))) \cdot (R(x) + (P(x))) = Q(x)R(x) + (P(x)),$$

or, in terms of the bracket notation for cosets,

$$[Q(x)] \cdot [R(x)] = [Q(x)R(x)]. \quad (10.15)$$

The fact that this multiplication is unambiguous follows from an argument analogous to that used in the last paragraph of the previous section (Exercise 10.3.24).

Thus, in $\mathbb{Z}_2[x]/(x^2 + x + 1)$,

$$[1 + x] \cdot [1 + x] = [(1 + x)^2] = [1 + 2x + x^2] = [x],$$

whereas in $\mathbb{Z}_2[x]/(x^2 + 1)$,

$$[1 + x] \cdot [1 + x] = [(1 + x)^2] = [1 + 2x + x^2] = [0].$$

One more concept is necessary for the formulation of this section's main theorem. If F is a field and E is a subset of F which also constitutes a field with respect to the arithmetical operations it inherits from F , then E is a *subfield* of F and F is an *extension* of E . Thus, the complex numbers are an extension of the reals which, in turn, are an extension of the rational numbers. Similarly, every Galois field $\text{GF}(p, P(x))$ is an extension of \mathbb{Z}_p . We shall use this terminology even when F only contains a subfield that is isomorphic to E rather than E itself. In general, isomorphic fields will be identified.

Theorem 10.16 If F is any field and $P(x)$ is any irreducible polynomial over F , then $F[x]/(P(x))$ is a field extension of F that contains a zero of $P(x)$.

Proof. We first demonstrate that $F[x]/(P(x))$ is indeed a field. Since the addition and multiplication of the elements of $F[x]/(P(x))$ are given by Equations 10.13 and 10.15 it follows that these operations satisfy all but the last of the requirements in Section 6.1 simply because the usual addition and multiplication of polynomials also satisfy those requirements. Thus, the additive and multiplicative identities of $F[x]/(P(x))$ are $[0]$ and $[1]$, and, by way of example, commutativity holds for multiplication in $F[x]/(P(x))$ because

$$[Q(x)] \cdot [R(x)] = [Q(x)R(x)] = [R(x)Q(x)] = [R(x)] \cdot [Q(x)].$$

However, since the multiplicative inverse of a polynomial is not a polynomial, more work is required to demonstrate that the nonzero elements of $F[x]/(P(x))$ possess multiplicative inverses. The argument we give here is a slight modification of the proof of Lemma 7.4

Let $Q(x)$ be any polynomial over F that is not in $(P(x))$, i.e., $Q(x)$ is a polynomial that is not divisible by $P(x)$. Since $P(x)$ is irreducible this is tantamount to saying that

$P(x)$ and $Q(x)$ are relatively prime, and hence there exist $A(x), B(x) \in F[x]$ such that

$$A(x)Q(x) + B(x)P(x) = 1.$$

Passing on to cosets, we conclude that

$$[A(x)] \cdot [Q(x)] = [A(x)Q(x)] = [1 - B(x)P(x)] = [1]$$

the justification for the last equality being that $1 - B(x)P(x)$ and 1 differ by an element of $(P(x))$ and hence they belong to the same coset of $(P(x))$. Since $[1]$ is the multiplicative identity of $F[x]/(P(x))$, it follows that $[A(x)]$ is the multiplicative inverse of $[Q(x)]$, and so $F[x]/(P(x))$ is indeed a field. It is clear from the definitions of addition and multiplication in $F[x]/(P(x))$ that the collection of cosets $F' = \{ [r] \mid r \in F \}$ constitutes a subfield of $F[x]/(P(x))$ that is isomorphic to F , the isomorphism $f: F \rightarrow F'$ being defined by $f(r) = [r]$. Hence, $F[x]/(P(x))$ is indeed an extension of F . Finally, we show that the coset $[x]$ is a zero of the polynomial $P(x)$ over the field $F[x]/(P(x))$. If

$$P(x) = a_0x^n + a_1x^{n-1} + \cdots + a_n$$

is a polynomial over F , then as a polynomial over $F[x]/(P(x))$, it should be written as

$$P(y) = [a_0]y^n + [a_1]y^{n-1} + \cdots + [a_n].$$

The reason the variable x of $P(x)$ needs to be replaced by y is that $[x]$ has become an element of the field $F[x]/(P(x))$. If the variable y of $P(y)$ is now replaced by this element $[x]$ of $F[x]/(P(x))$, then

$$\begin{aligned} P([x]) &= [a_0] \cdot [x]^n + [a_1] \cdot [x]^{n-1} + \cdots + [a_n] \\ &= [a_0x^n] + [a_1x^{n-1}] + \cdots + [a_n] \\ &= [a_0x^n + a_1x^{n-1} + \cdots + a_n] \\ &= [P(x)] = [0]. \end{aligned}$$

In other words, the element $[x]$ of $F[x]/(P(x))$ is a zero of the polynomial $P(y)$, which, of course, is identical with the polynomial $P(x)$. ■

Amongst other things, this theorem justifies Galois's assertion (Section 7.1) that every polynomial that is irreducible over \mathbb{Z}_p has a zero which we called its Galois imaginary. Being mortals ourselves, we will not comment on why Galois could make such an assertion without falling into the pits where most unjustified assumptions lead.

Before this theorem is applied to the creation of some new fields, we will show how it can be used to give a rigorous construction of the complex numbers. The polynomial $x^2 + 1$ is irreducible over \mathbb{R} and hence, by the above theorem, $\mathbb{R}[x]/(x^2 + 1)$ is a field that contains a subfield isomorphic to \mathbb{R} and in which $[x]$ satisfies the equation

$$[x]^2 + [1] = [x^2 + 1] = [0].$$

Since the function $f(r) = [r]$ is an isomorphism of \mathbb{R} onto a subfield of $\mathbb{R}[x]/(x^2 + 1)$, we may identify $[r]$ with r . In addition, let us label $[x]$ by i , so that the typical element of $\mathbb{R}[x]/(x^2 + 1)$ is

$$[a + bx] = [a] + [b] \cdot [x] = a + bi$$

where $a, b \in \mathbb{R}$ and

$$i^2 = [x]^2 = -[1] = -1.$$

It is now clear that $\mathbb{R}[x]/(x^2 + 1)$ is in fact isomorphic to the complex number system. This might be the place to note that this particular application is in fact the origin of Theorem 10.16. In his paper of 1847, Cauchy used this very approach to justify the existence of complex numbers.

The Galois fields of Chapter 7 can also be regarded as a special case of this construction. If $P(x)$ is an irreducible polynomial of degree ν over \mathbb{Z}_p , then, according to Proposition 10.14, the elements of the field $\mathbb{Z}_p[x]/(P(x))$ all have the form

$$a_0 + a_1[x] + a_2[x]^2 + \cdots + a_{\nu-1}[x]^{\nu-1}$$

which, when $[x]$ is interpreted as the Galois imaginary i , is identical with Expression 7.2. When this identification between the elements of $\mathbb{Z}_p[x]/(P(x))$ and $\text{GF}(p, P(x))$ is carried out, the arithmetical operations of the one are also identical with those of the other. In other words, $\mathbb{Z}_p[x]/(P(x))$ and $\text{GF}(p, P(x))$ are isomorphic fields. This observation is a special case of the following very general and very strong theorem whose proof falls outside the scope of this text. The order of a field is the number of elements it contains.

Theorem 10.17 Every finite field has order p^n for some prime p and some positive integer n . Given any such p and n there is a field of order p^n , and any two fields of the same finite order are isomorphic.

Consider next the field $F = \text{GF}(2, x^2 + x + 1)$ where α is the Galois imaginary associated with $x^2 + x + 1$. That is, $F = \{0, 1, \alpha, 1 + \alpha\}$ where $\alpha^2 = \alpha + 1$. The quadratic polynomial $x^2 + x + \alpha$ has no zeroes in F and is therefore irreducible over it. In accordance with Theorem 10.16

$$F[x]/(x^2 + x + \alpha)$$

is a field. If we write λ for $[x]$ and replace $[r]$ by r for all $r \in F$, the elements of this field are

$[0] = 0,$	$[1] = 1,$
$[\alpha] = \alpha,$	$[1 + \alpha] = 1 + \alpha,$
$[x] = \lambda,$	$[1 + x] = 1 + \lambda,$
$[\alpha + x] = \alpha + \lambda,$	$[1 + \alpha + x] = 1 + \alpha + \lambda,$
$[\alpha x] = \alpha\lambda,$	$[1 + \alpha x] = 1 + \alpha\lambda,$
$[\alpha + \alpha x] = \alpha + \alpha\lambda,$	$[1 + \alpha + \alpha x] = 1 + \alpha + \alpha\lambda,$
$[(1 + \alpha)x] = \lambda + \alpha\lambda,$	$[1 + (1 + \alpha)x] = 1 + \lambda + \alpha\lambda,$
$[\alpha + (1 + \alpha)x] = \alpha + \lambda + \alpha\lambda,$	$[1 + \alpha + (1 + \alpha)x] = 1 + \alpha + \lambda + \alpha\lambda.$

The elements of this field are to be added and multiplied as indicated by the following examples:

$$(1 + \alpha\lambda) + (1 + \alpha + \lambda + \alpha\lambda) = 2 + \alpha + \lambda + 2\alpha\lambda = \alpha + \lambda,$$

and, bearing in mind that $\alpha^2 = \alpha + 1$ and $\lambda^2 = \lambda + \alpha$,

$$\begin{aligned}
 (1 + \alpha\lambda)(1 + \alpha + \lambda + \alpha\lambda) &= 1 + \alpha + \lambda + \alpha\lambda + \alpha\lambda + \alpha^2\lambda + \alpha\lambda^2 + \alpha^2\lambda^2 \\
 &= 1 + \alpha + \lambda + (1 + \alpha)\lambda + \alpha(\lambda + \alpha) + (1 + \alpha)(\lambda + \alpha) \\
 &= 1 + \alpha + \lambda + \lambda + \alpha\lambda + \alpha\lambda + \alpha^2 + \lambda + \alpha + \alpha\lambda + \alpha^2 \\
 &= 1 + \lambda + \alpha\lambda.
 \end{aligned}$$

Since this field has the same number of elements as $\text{GF}(2, x^4 + x + 1)$, it follows from the unproved Theorem 10.17 that they should be isomorphic. This is borne out by the computation

$$\lambda^4 = (\lambda^2)^2 = (\lambda + \alpha)^2 = \lambda^2 + \alpha^2 = \lambda + \alpha + \alpha + 1 = \lambda + 1,$$

which shows that λ is a zero of the polynomial $x^4 + x + 1$ over \mathbb{Z}_2 . In other words, we can think of λ as the Galois imaginary of the field $\text{GF}(2, x^4 + x + 1)$.

Let us examine some extensions of the rationals \mathbb{Q} . Since there is no rational number whose square equals 2, it follows that the polynomial $x^2 - 2$ is irreducible over \mathbb{Q} . Consequently, Theorem 10.16 yields

$$\mathbb{Q}[x]/(x^2 - 2)$$

as a new field. In accordance with Proposition 10.14, the elements of this field all have the form $[a] + [b][x]$ where a and b are rational numbers and $[x]$ is a quantity such that

$$[x]^2 - [2] = [x^2 - 2] = [0].$$

Since $[x]$ behaves just like a square root of 2, we denote it by the formal symbol $\sqrt{2}$. This formalism notwithstanding, note that

$$(\sqrt{2})^2 = [x]^2 = [2] = 2$$

in this field. If we persist in identifying $[r]$ with r for each rational number r , then the elements of this new field have the form $a + b\sqrt{2}$ for $a, b \in \mathbb{Q}$.

The addition and multiplication of the elements of this field are given by

$$(a_1 + b_1\sqrt{2}) + (a_2 + b_2\sqrt{2}) = (a_1 + a_2) + (b_1 + b_2)\sqrt{2}$$

and

$$(a_1 + b_1\sqrt{2})(a_2 + b_2\sqrt{2}) = (a_1a_2 + 2b_1b_2) + (a_1b_2 + a_2b_1)\sqrt{2}.$$

Note that

$$(a + b\sqrt{2})^{-1} = \frac{a}{a^2 - 2b^2} + \frac{-b}{a^2 - 2b^2}\sqrt{2}.$$

Let us denote this new field by $\mathbb{Q}[x]/(x^2-2)$ by F_1 . The field F_1 can itself serve as the ground field for the construction of another field. Consider, for example, the polynomial x^2-3 . This polynomial is irreducible over F_1 . To justify this claim it suffices to show that the equation

$$(x + y\sqrt{2})^2 - 3 = 0$$

has no solution wherein both x and y are rational. However, this equation simplifies to

$$x^2 + 2y^2 + 2xy\sqrt{2} = 3$$

or

$$\sqrt{2} = \frac{3 - x^2 - 2y^2}{2xy}$$

which cannot have rational solutions since 2 is known not to be a rational number.

The quadratic x^2-3 being irreducible over F_1 , it yields yet a new field

$$F_1[x]/(x^2-3)$$

whose typical element, when $[x]$ is symbolized by $\sqrt{3}$, is

$$(a_1 + b_1\sqrt{2}) + (a_2 + b_2\sqrt{2})\sqrt{3}.$$

If we abbreviate $\sqrt{2}\sqrt{3}$ to the symbol $\sqrt{6}$, then all the elements of $F_1[x]/(x^2-3)$ can be written in the form

$$a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6},$$

with $a, b, c, d \in \mathbb{Q}$.

It is clear that Theorem 10.16 can be used to construct a myriad of new fields. The general theory of these fields and their classification falls outside the scope of this book.

We conclude this section with a warning about a possible source of confusion. While the powerful Theorem 10.16 resembles the Fundamental Theorem of Algebra, which asserts the existence of complex zeroes to every complex polynomial, the two theorems are distinct. The zeroes whose existence is guaranteed by Theorem 10.16 need not belong to the ground field, as is exemplified by the polynomial $x^2 + x + 1$ over \mathbb{Z}_2 . The Fundamental Theorem of Algebra, above and beyond asserting the mere existence of zeroes of complex polynomials also places them back in the ground field, which Theorem 10.16 does not do.

Exercises 10.3

For each pair F and $P(x)$ in Exercises 10.3.1 to 10.3.8, describe a subfield of the complex numbers that is isomorphic to $F[x]/(P(x))$.

1. $F = \mathbb{Q}, P(x) = x^2 - 5$
2. $F = \mathbb{Q}, P(x) = x^3 - 2$
3. $F = \mathbb{Q}, P(x) = x^2 + 1$
4. $F = \mathbb{Q}, P(x) = x^2 + 25$
5. $F = \mathbb{Q}, P(x) = x^2 + x + 1$
6. $F = \mathbb{Q}[x]/(x^2 - 2), P(x) = x^2 - 5$
7. $F = \mathbb{Q}[x]/(x^2 - 2), P(x) = x^2 + 1$
8. $F = \mathbb{R}, P(x) = x^2 + 5$

For each of the fields in Exercises 10.3.9 to 10.3.12, find a field that is described in Chapter 7 and is isomorphic to it.

9. $\mathbb{Z}_2[x]/(x^2 + x + 1)$
10. $\mathbb{Z}_2[x]/(x^3 + x^2 + 1)$
11. $\mathbb{Z}_2[x]/(x^4 + x^3 + x^2 + x + 1)$
12. $\mathbb{Z}_3[x]/(x^2 + x + 2)$

Explain why each of the fields in Exercises 10.3.13 to 10.3.16 is finite.

13. $\mathbb{Z}_2[x]/(x^2 + x + 1)$
14. $\mathbb{Z}_2[x]/(x^3 + x + 1)$
15. $\mathbb{Z}_2[x]/(x^4 + x^3 + x^2 + x + 1)$
16. $\mathbb{Z}_5[x]/(x^2 + 4x + 2)$

17. Prove that the set of real numbers

$$\{a + b\sqrt{2} + c\sqrt{3} \mid a, b, c \in \mathbb{Q}\}$$

does not constitute a subfield of the real numbers (with respect to the usual arithmetical operations).

18. Prove that the set of real numbers

$$\{a + b\sqrt{7} + c\sqrt{11} \mid a, b, c \in \mathbb{Q}\}$$

does not constitute a subfield of the real numbers (with respect to the usual arithmetical operations).

19. What is the multiplicative inverse of $2 + 3[x]$ in the field $\mathbb{Q}[x]/(x^2 - 5)$?
20. What is the multiplicative inverse of $3 + 2[x]$ in the field $\mathbb{Q}[x]/(x^2 - 7)$?
21. What is the multiplicative inverse of the element $1 + [x]$ in the field $\mathbb{Q}[x]/(x^2 + 5)$?

22. Let r be any rational number such that r is not rational. Find a formula for the multiplicative inverse of $a + b[x]$ in $\mathbb{Q}[x]/(x^2 - r)$.
23. Suppose $P(x)$ is a polynomial of degree d over a field F . Prove that there exists an extension F' of F such that, counting multiplicities, $P(x)$ has d zeroes in F' .
24. Prove that the multiplication of the elements of $F[x]/(P(x))$ defined in Equation 10.15 is unambiguous.
25. Show that if $P(x)$ is a reducible element of $F[x]$, then the multiplication defined in Equation 10.15 does not yield a field.

10.4 Galois Groups and the Resolvability of Equations

We would like to conclude this chapter with a brief account of how the young Galois settled the question of the algebraic resolvability of equations. Because of the introductory nature of this text, such an account must of necessity be superficial, and a more complete exposition of the theory can be found in many graduate texts.

Briefly put, if $P(x)$ is an irreducible polynomial, with either real or complex coefficients, then Galois associated a certain group of permutations with the equation $P(x) = 0$ and then proved that an appropriate analysis of the group yields the answer as to whether or not this equation is algebraically resolvable. We shall now discuss both the group and its analysis.

Let $P(x)$ be an irreducible polynomial with integer coefficients. The Galois group of the polynomial equation $P(x) = 0$ is a group of permutations of the roots of the equation (recall that the existence of these roots is guaranteed by the Fundamental Theorem of Algebra of Section 3.3) that enjoys two properties:

First, every rational expression in the roots that is invariant under all the permutations in the group has a rational expression in the coefficients of the equation.

Second, conversely, every rational expression in the roots that is also a rational expression in the coefficients is necessarily invariant under all the permutations of the group.

Consider, for example, the cyclotomic equation $x^4 + x^3 + x^2 + x + 1 = 0$. Since $x^5 - 1 = x^4 + x^3 + x^2 + x + 1$, it follows that the roots of this equation are ε , ε^2 , ε^3 , and ε^4 , where ε is the first 5th root of unity. It is known that the Galois group of this equation is the permutation group of order 4 generated by the cycle $\sigma = (\varepsilon \varepsilon^2 \varepsilon^4 \varepsilon^3)$. We will now illustrate the meaning of the two properties above by investigating this group's effect on some rational expressions in these roots. Suppose first that a , b , c , and d are

integers such that the expression

$$\varphi = (\varepsilon)^a (\varepsilon^2)^b (\varepsilon^4)^c (\varepsilon^3)^d$$

is invariant under σ (and so it is necessarily also invariant under all the permutations in the Galois group since, in this case, they are all powers of σ). This means that

$$(\varepsilon)^a (\varepsilon^2)^b (\varepsilon^4)^c (\varepsilon^3)^d = (\varepsilon^2)^a (\varepsilon^4)^b (\varepsilon^3)^c (\varepsilon)^d \quad (10.18)$$

and hence

$$a + 2b + 4c + 3d \equiv 2a + 4b + 3c + d \pmod{5}, \quad (10.19)$$

and subtraction surprisingly yields

$$0 \equiv a + 2b + 4c + 3d \pmod{5}. \quad (10.20)$$

From this it immediately follows that

$$\varphi = \varepsilon^{a+2b+4c+3d} = 1.$$

Thus, as required by the first property, the invariance of φ under the permutations of the Galois group was sufficient to guarantee its rationality. Suppose now that we only know φ to be rational. Since φ is necessarily a fifth root of unity, it follows that $\varphi = 1$ and so Equation 10.20 holds. Equation 10.20 entails Equation 10.19 and this one implies Equation 10.18. In other words, from the mere assumption of the rationality of φ it was possible to prove its invariance under σ and all the elements of the Galois group.

Galois himself gives two examples of these groups. If x_1, x_2, \dots, x_n denote the roots of the general equation

$$x^n + a_1 x^{n-1} + \dots + a_{n-1} x + a_n = 0,$$

then the Galois group of this equation consists of all the permutations of these roots. The Fundamental Theorem of Symmetric Polynomials, briefly mentioned in Section 6.4, asserts that every symmetric rational polynomial of the variables x_1, x_2, \dots, x_n can be expressed as a function of the elementary symmetric polynomials of these variables. By Theorem 6.19, when these variables denote the roots of the above equation, their elementary symmetric functions equal $(-1)^k a_k$ for $k = 1, 2, \dots, n$. Thus, the first condition is

satisfied. That the second condition is satisfied is harder to show, and we will not do so here.

The second example Galois gives is that of the cyclotomic equation $x^p - 1 = 0$ where p is a prime number. Since the polynomial $x^n - 1$ is never irreducible, it is necessary to divide out the factor $x - 1$ after which we get

$$x^{p-1} + x^{p-2} + \cdots + x + 1 = 0 \quad (10.21)$$

which can be proved to be irreducible (for prime p). The Galois group of this equation is the cyclic group $\langle \sigma \rangle$ where

$$\sigma = (\zeta \zeta^k \zeta^{k^2} \cdots \zeta^{k^{p-2}})$$

where ζ is any primitive p -th root of unity and k is any primitive root modulo p . In the special case of $p = 17$ where k is taken to be 3, we get

$$\sigma = (\zeta \zeta^3 \zeta^9 \zeta^{10} \zeta^{13} \zeta^5 \zeta^{15} \zeta^{11} \zeta^{16} \zeta^{14} \zeta^8 \zeta^7 \zeta^4 \zeta^{12} \zeta^2 \zeta^6).$$

Since σ is in general a cyclic permutation of $p - 1$ elements, it follows that the Galois group of Equation 10.21 is isomorphic to $(\mathbb{Z}_{p-1}, +)$.

The analysis of the Galois group that determines the algebraic resolvability of its originating equation has the form of a recursive procedure. If the group G has prime order p (so that it is isomorphic to $(\mathbb{Z}_p, +)$), then the equation is algebraically resolvable. Next, if G has composite order and it contains no proper normal subgroups, then the originating equation is not algebraically resolvable. If neither of these conditions hold, then G is a group of a composite order with a proper normal subgroup, say H . Now apply the same analysis to both H and G/H . If at any time we encounter a group of composite order without a proper normal subgroup, then the originating equation is not algebraically resolvable. Since at each stage the orders of H and G/H are smaller than the order of G , this process is bound to terminate. If all the groups of composite order encountered in this procedure have proper normal subgroups, the originating equation is algebraically resolvable. Otherwise, it is not so resolvable.

As an example we consider the cyclotomic equation $x^{13} - 1 = 0$, whose irreducible part is

$$x^{12} + x^{11} + x^{10} + \cdots + x + 1 = 0 \quad (10.22)$$

and whose Galois group is isomorphic to $(\mathbb{Z}_{12}, +)$. The order of $(\mathbb{Z}_{12}, +)$ is composite, and $H = \{0, 6\}$ is a subgroup of H . Since $(\mathbb{Z}_{12}, +)$ is an abelian group, H is necessarily normal, and since $(\mathbb{Z}_{12}, +)$ is cyclic, so is $G_1 = (\mathbb{Z}_{12}, +)/H$. Now H has order 2 which is a prime and so we are done with it. On the other hand, G_1 has the composite order 6, and so, by Theorem 9.14, G_1 is isomorphic to $(\mathbb{Z}_6, +)$. Consequently, G_1 itself has a (necessarily normal) subgroup G_2 of order 3. Thus G_2 has order 3 and G_1/G_2 has order 2, both being prime numbers. It follows from Galois's theory that Equation 10.22 is indeed algebraically resolvable, whose resolution was first investigated by Gauss.

Let us examine the general quintic equation. As noted above, its Galois group is isomorphic to S_5 which has the composite order $5! = 120$. The symmetric group S_5 has the group A_5 of all the even permutations of $\{1, 2, 3, 4, 5\}$ as a subgroup, and it follows from Proposition 10.3 that A_5 is a normal subgroup of S_5 such that S_5/A_5 has order 2. However, while A_5 has a plenitude of proper subgroups, we will now show that none of these subgroups is normal. Consequently, the general quintic equation is not resolvable by radicals, a fact that had of course already been proved by Abel.

Proposition 10.23 The group A_5 has no proper normal subgroups.

Proof. Suppose H is a normal subgroup of A_5 that contains some nonidentity element. We will show that H necessarily equals A_5 by demonstrating the following two statements:

- H contains a 3-cycle;
- H contains all the 3-cycles of A_5 .

Since we already know that every even permutation is expressible as the composition of 3-cycles (Exercise 8.4.13), it will follow that $H = A_5$.

Proof that H contains a 3-cycle: Since all the elements of A_5 are even permutations, it follows that their disjoint cycle decompositions consist either of a single 5-cycle, a single 3-cycle, or a pair of transpositions such as $(1\ 2)(3\ 4)$. If H contains a 5-cycle, say $(1\ 2\ 3\ 4\ 5)$, then, because H is a normal subgroup of A_5 , it must also contain the element

$$\begin{aligned} (1\ 2\ 3\ 4\ 5)^{-1}[(1\ 2\ 3)(1\ 2\ 3\ 4\ 5)(1\ 2\ 3)^{-1}] \\ = (5\ 4\ 3\ 2\ 1)(1\ 2\ 3)(1\ 2\ 3\ 4\ 5)(3\ 2\ 1) = (1\ 3\ 5). \end{aligned}$$

If H contains a pair of transpositions, say $(1\ 2)(3\ 4)$, then, because H is a normal subgroup, it must also contain the element

$$\begin{aligned} & (1\ 2)(3\ 4)[(2\ 5)(3\ 4)](1\ 2)(3\ 4)[(2\ 5)(3\ 4)]^{-1} \\ &= (1\ 2)(3\ 4)(2\ 5)(3\ 4)(1\ 2)(3\ 4)(3\ 4)(2\ 5) = (1\ 5\ 2). \end{aligned}$$

Thus, if H contains any nonidentity element, then it necessarily contains some 3-cycle.

Proof that H contains all the 3-cycles of A_5 : By the above we know that H contains some 3-cycle, say $(1\ 2\ 3)$. Moreover, if (abc) is any 3-cycle of A_5 , then by Exercise 8.2.23

$$(1\ 2\ 3\ 4\ 5)(abc)(1\ 2\ 3\ 4\ 5)^{-1} = (a+1\ b+1\ c+1)$$

where the addition is computed modulo 5. Hence, since $(1\ 2\ 3\ 4\ 5) \in A_5$ and H is normal in A_5 , it follows that H must contain the 3-cycles $(1\ 2\ 3)$, $(2\ 3\ 4)$, $(3\ 4\ 5)$, $(4\ 5\ 1)$, and $(5\ 1\ 2)$ as well as their inverses. Since

$$[(1\ 2)(3\ 4)](1\ 2\ 3)[(1\ 2)(3\ 4)]^{-1} = (1\ 2)(3\ 4)(1\ 2\ 3)(3\ 4)(1\ 2) = (1\ 4\ 2),$$

it follows for similar reasons that H must also contain the 3-cycles $(1\ 4\ 2)$, $(2\ 5\ 3)$, $(3\ 1\ 4)$, $(4\ 2\ 5)$, and $(5\ 3\ 1)$ and their inverses. As these exhaust all the twenty 3-cycles of A_5 , the proof is complete. ■

The above proposition generalizes to the statement that A_n is simple for each $n \geq 5$.

This concludes our attempt at an elementary description of Galois theory. It remains only to say a few words about the subsequent evolution of group theory.

We saw that groups caught the attention of mathematicians because they provided the key to the question of which polynomials equations are algebraically solvable. Thus, the Galois group of an equation contains all the information that is required to decide on its algebraic resolvability. Mathematicians subsequently went on to try to classify all these new structures and eventually a very clear cut and apparently also very difficult question crystallized. Is every finite group necessarily isomorphic to the Galois group of some equation with integer coefficients? As of the writing of this text the answer to this question is still unknown.

Groups that contain no proper normal subgroups were seen to play a key role in Galois theory and are called simple groups. The commutative simple groups are the cyclic groups

of prime order. The group A_5 , which is the subject of Proposition 10.23, is the smallest of the noncommutative simple groups. The appellation “simple” is not to be taken literally. These groups have in fact very complicated structures. The joint efforts of hundreds of group theorists resulted recently in the complete classification of the finite simple groups. This monumental work occupies about 14,000 pages of mathematical publications. The last finite simple group to be identified is affectionately known as the monster. Its order is

$$2^{46} \cdot 3^{20} \cdot 5^9 \cdot 7^6 \cdot 11^2 \cdot 13^2 \cdot 17 \cdot 19 \cdot 23 \cdot 29 \cdot 31 \cdot 41 \cdot 47 \cdot 59 \cdot 71 \approx 10^{54}.$$

The monster happens to be the group of vertex symmetries of a solid that resides in a space of 196,883 dimensions. Strangely enough, the number 196,884 features in some applications of non-Euclidean geometry to number theory, but that’s another story.

Exercises 10.4

1. Prove that no commutative group of composite (or infinite) order is simple.
2. Suppose G is a subgroup of S_n . Prove that, if G is simple and $o(G) > 2$, then $G \subset A_n$.
3. Prove that the group of vertex symmetries of the regular octahedron (Figure 9.6) is not simple.
4. Prove that the group of vertex symmetries of the cube (Figure 9.5) is not simple.
5. Suppose that G is a finite simple group and $1_G \neq a \in G$. Prove that every element of G can be expressed as a product of elements of the conjugacy class $C(a)$ (cf. Exercises 10.1.24 to 10.1.30).
6. Prove that A_6 is not a simple group.

Chapter Summary

We have shown that new groups can be obtained from old ones by the quotient operation. This operation was applied to additive groups of polynomials to produce a host of new and old fields. Finally, the notion of quotient groups permitted us to formulate Galois’s criterion for the resolvability of algebraic equations.

Chapter Review Exercises

Mark the following true or false.

1. If K is the Klein 4-group, then KK contains more elements than K .
2. K is a normal subgroup of D_4 .
3. \mathbb{Z}_{59}^* has a subgroup that is isomorphic to K .
4. The number of distinct elements of $\mathbb{Z}_5[x]/(x^2 + 2x + 1)$ is 25.
5. $(\mathbb{Z}_5, +)$ is a simple group.
6. The complex numbers constitute an extension of the rational numbers.

New Terms

conjugacy class, 234	monomorphism, 235
conjugate, 234	normal subgroup, 228
extension, 246	quotient group, 230
homomorphism, 234	signature, 236
kernel, 237	subfield, 246

Supplementary Exercises

1. Write a computer script which will decide whether a given subgroup of some group is normal. If the answer is yes, write out a multiplication table for the quotient group.
2. Prove that the alternating group A_n is simple for $n \geq 5$.

Chapter 11



TOPICS IN ELEMENTARY GROUP THEORY

THIS CHAPTER DISPLAYS some of the methods and results of elementary group theory. Specifically, we demonstrate how many more groups can be constructed and classify, up to isomorphism, all the groups of orders $2p$ and p^2 for p prime.

11.1 The Direct Product of Groups

In this section we describe one of many methods for combining groups to produce new groups. If G and H are two groups, then their *direct product*, denoted by $G \times H$, has as its elements the set of all the ordered pairs (g, h) where $g \in G$ and $h \in H$. The binary operation of $G \times H$ is defined by

$$(g, h)(g', h') = (g g', h h').$$

The associativity of this operation follows directly from the associativity of the group operations of G and H . The identity element of $G \times H$ is $(1_G, 1_H)$, and the inverse of (g, h) is the pair (g^{-1}, h^{-1}) . Thus, $\mathbb{Z}_2 \times \mathbb{Z}_2$ consists of the four pairs $(0, 0)$, $(1, 0)$, $(0, 1)$, and $(1, 1)$ where

$$(1, 1)(1, 0) = (1 + 1, 1 + 0) = (0, 1) \quad \text{and}$$

$$(1, 1)(1, 1) = (1 + 1, 1 + 1) = (0, 0).$$

Similarly, $\mathbb{Z}_2 \times \mathbb{Z}_3$ consists of the six pairs $(0, 0)$, $(0, 1)$, $(0, 2)$, $(1, 0)$, $(1, 1)$, and $(1, 2)$ where

$$(1, 1)(0, 2) = (1 + 0, 1 + 2) = (1, 0) \quad \text{and}$$

$$(1, 2)(0, 2) = (1 + 0, 2 + 2) = (1, 1).$$

It is clear that if G and H are finite groups, then $o(G \times H) = o(G)o(H)$. It is also easy to see that the function $f((g, h)) = (h, g)$ defines an isomorphism of $G \times H$ and $H \times G$. It is equally clear that the subgroups

$$G' = \{(g, 1_H) \mid g \in G\} \quad \text{and} \quad H' = \{(1_G, h) \mid h \in H\}$$

of $G \times H$ are isomorphic to G and H , respectively. In particular, $o((g, 1_H)) = o(g)$ and $o((1_G, h)) = o(h)$. The next proposition tells us how to determine the order of any element of $G \times H$.

Proposition 11.1 If g and h are elements of finite orders in the groups G and H , respectively, then $o((g, h))$ is the least common multiple of $o(g)$ and $o(h)$.

Proof. Let k be the least common multiple of $o(g)$ and $o(h)$ and let $d = o((g, h))$. Since

$$(g, h)^k = (g^k, h^k) = (1_G, 1_H)$$

it follows that d divides k . Conversely, since

$$(1_G, 1_H) = 1_{G \times H} = (g, h)^d = (g^d, h^d)$$

it follows that $1_G = g^d$ and $1_H = h^d$. Thus d is divisible by both $o(g)$ and $o(h)$, and so, by Exercise 4.2.31, d is divisible by k . Hence, $d = k$. ■

It follows from Proposition 11.1 that $\mathbb{Z}_2 \times \mathbb{Z}_2$ is a group of order 4 in which every nonidentity element has order 2, and hence this group is isomorphic to K . Similarly, the element $(1, 2)$ of $\mathbb{Z}_2 \times \mathbb{Z}_3$ has order $2 \cdot 3 = 6$, and hence this group is isomorphic to $(\mathbb{Z}_6, +)$. On the other hand, the group $\mathbb{Z}_2 \times \mathbb{Z}_4$ has order 8, is commutative, contains no elements of order 8, and contains an element of order 4, namely $(0, 1)$. This is enough information to justify the assertion that $\mathbb{Z}_2 \times \mathbb{Z}_4$ is not isomorphic to any of the previously encountered groups of order 8, namely, the Quaternion group, D_4 , $(\mathbb{Z}_8, +)$, and $\mathbb{Z}_2[x, \leq 2]$.

The preceding paragraph makes it clear that it would be useful to have some criteria for recognizing when a group is isomorphic to the direct product of some two other groups. This is now provided.

Proposition 11.2 Suppose the finite group P contains two normal subgroups G and H such that $G \cap H = \{1_P\}$ and $o(P) = o(G)o(H)$. Then $P \cong G \times H$.

Proof. We begin by proving that the elements of G commute with elements of H . Thus, suppose $g \in G$ and $h \in H$. Since G and H are normal in P it follows that

$$\begin{aligned} ghg^{-1}h^{-1} &= (ghg^{-1})h^{-1} \in HH = H \quad \text{and} \\ ghg^{-1}h^{-1} &= g(hg^{-1}h^{-1}) \in GG = G. \end{aligned}$$

Since $G \cap H = \{1_P\}$, it follows that $ghg^{-1}h^{-1} = 1_P$ and hence $gh = hg$ for all $g \in G$ and $h \in H$.

We are now ready to prove the required isomorphism. Let $f((g, h)) = gh$. This is clearly a function from $G \times H$ into P . If $f((g, h)) = f((g', h'))$ then $gh = g'h'$ or $(g')^{-1}g = h'h^{-1}$. However, $(g')^{-1}g \in G$ and $h'h^{-1} \in H$, and hence, since $G \cap H = \{1_P\}$, it follows that either $(g')^{-1}g = h'h^{-1} = 1_P$ or both $g = g'$ and $h = h'$. It follows that f maps distinct elements of $G \times H$ to distinct elements of P . Since $o(P) = o(G)o(H) = o(G \times H)$, it follows that f does indeed match all the elements of $G \times H$ with those of P . Finally, making use of the above-proved commutativity, note that

$$f((g, h)(g', h')) = f((gg', hh')) = gg'h'h' = ghg'h' = f((g, h))f((g', h')),$$

so that f is indeed an isomorphism. ■

As a consequence of Proposition 11.2 we show that if p and q are any two distinct prime numbers, then $(\mathbb{Z}_{pq}, +) \cong (\mathbb{Z}_p, +) \times (\mathbb{Z}_q, +)$. Let $P = (\mathbb{Z}_{pq}, +)$, $G = \langle p \rangle \cong (\mathbb{Z}_q, +)$, and $H = \langle q \rangle \cong (\mathbb{Z}_p, +)$. Then P , G , and H satisfy the hypotheses of Proposition 11.2, and hence the desired conclusion follows. The next corollary illustrates a somewhat more complicated application of Proposition 11.2. First, however, we note that it is clear that for any three groups G , H , and J that $(G \times H) \times J \cong G \times (H \times J)$, this isomorphism being established by the function $f(((g, h), k)) = (g, (h, k))$. Consequently, we can unambiguously write $G \times H \times J$ for $(G \times H) \times J$ and $G \times (H \times J)$. Further, if k is any positive integer, then we denote the direct product of k copies of G by G^k .

Corollary 11.3 If P is a finite group in which every nonidentity element has order 2, then there exists a nonnegative integer k such that $P \cong (\mathbb{Z}_2, +)^k$.

Proof. We proceed by induction on the order of P , the conclusion being trivially valid when $o(P) \leq 2$. We therefore assume that $o(P) = n$ and that the proposition holds for all groups of order less than n . Let H be a proper subgroup of P which is contained in

no other proper subgroup of P and let a be any element of P that is not in H . Then, because P is a commutative group (Exercise 9.2.29), $H' = H \cup (aH)$ is also a subgroup of P . Since H' contains both H and a , it follows from the maximality of H that $H' = P$. Thus H has index 2 in P and so $o(P) = 2o(H) = o(\langle a \rangle)o(H)$. It now follows from Proposition 11.2 that $P \cong \langle a \rangle \times H \cong (\mathbb{Z}_2, +) \times H$. Since every element of H also has order 2 and H has order less than n , it follows from the induction hypothesis that $H \cong (\mathbb{Z}_2, +)^k$ for some k , and hence $P \cong (\mathbb{Z}_2, +) \times (\mathbb{Z}_2, +)^k = (\mathbb{Z}_2, +)^{k+1}$. ■

Exercises 11.1

1. Let G and H be finite groups. Prove that the sets $G' = \{(g, 1_H) \mid g \in G\}$ and $H' = \{(1_G, h) \mid h \in H\}$ are normal subgroups of $G \times H$. Prove that $(G \times H)/G' \cong H$ and $(G \times H)/H' \cong G$.
2. Let p be a prime and let G be a finite commutative group in which every nonidentity element has order p . Prove that $G \cong (\mathbb{Z}_p, +)^k$ for some nonnegative integer k .
3. Let p be a prime. Prove that every commutative group of order p^2 is isomorphic to either $(\mathbb{Z}_{p^2}, +)$ or $(\mathbb{Z}_p, +)^2$.
4. Suppose m and n are relatively prime positive integers. Prove that $(\mathbb{Z}_{mn}, +) \cong (\mathbb{Z}_m, +) \times (\mathbb{Z}_n, +)$.
5. Prove that $(\mathbb{Z}_{209}, +) \times (\mathbb{Z}_{221}, +) \cong (\mathbb{Z}_{143}, +) \times (\mathbb{Z}_{323}, +)$.

A group is said to be a *decomposable group* if it is isomorphic to the direct product of two nontrivial groups. Otherwise, it is *indecomposable*.

6. Prove that $(\mathbb{Z}_n, +)$ is indecomposable if and only if $n = p^r$ for some prime p .
7. Prove that D_5 is indecomposable.
8. Prove that S_3 is indecomposable.
9. Prove that S_4 is indecomposable.
10. Prove that S_5 is indecomposable.
11. Prove that the Quaternion group is indecomposable.
12. Prove that if $1 \leq k \leq n$, then S_n contains a subgroup $P \cong S_k \times S_{n-k}$.
13. Prove that if G and H are finite groups then $G \times H$ is cyclic if and only if both G and H are cyclic and their orders are relatively prime.

14. Let p be any prime. Prove that for every positive integer n there are at least n nonisomorphic groups of order p^n .
15. Prove that every finite group G of order greater than 2 has an automorphism that is distinct from the identity function.

11.2 More Classifications

We begin by reviewing some information that was relegated to the exercises of Chapters 9 and 10. Two elements a and b of a group G are said to be *conjugate* if there exists an element x such that $xax^{-1} = b$. The set of all the elements of G that are conjugate to a is denoted by $C(a)$ and is called the *conjugacy class* of a . Note that if a is conjugate to b , then

$$a = x^{-1}xax^{-1}x = x^{-1}bx = x^{-1}b(x^{-1})^{-1}$$

so that b is also conjugate to a . Moreover, if b is also conjugate to $c \in G$, say $b = ycy^{-1}$, then

$$c = y^{-1}by = y^{-1}(xax^{-1})y = (y^{-1}x)a(y^{-1}x)^{-1},$$

so that a is also conjugate to c . These observations have the following consequence.

Lemma 11.4 If a and b are elements of the group G , then $C(a)$ and $C(b)$ are either identical or disjoint.

Proof. Suppose $C(a)$ and $C(b)$ are not disjoint, so that $c \in C(a) \cap C(b)$ for some $c \in G$. It then follows from the above observations that a and b , both being conjugate to c , are also conjugate to each other. However, if $d \in C(a)$, then d , being conjugate to a , is also conjugate to b , so that $C(a) \subset C(b)$. By symmetry, $C(b) \subset C(a)$, and hence $C(a) = C(b)$. ■

The *centralizer* Z_a of the element a of the group G consists of all the elements x in G such that $xa = ax$ (see Exercises 9.4.43 to 9.4.51). Note that Z_a always contains 1_G , a , and every power of a . If a commutes with every element of G , so that $Z_a = G$, then $xax^{-1} = a$ for all $x \in G$, so that $C(a) = a$. If G is the Quaternion group (Table 9.3) $Z_1 = Z_d = G$ and $Z_x = \{1, d, x, x^{-1}\}$ for all other $x \in G$. The following proposition shows that there is a very strong relationship between centralizers and conjugacy classes.

Proposition 11.5 Let G be a group and $a \in G$. Then Z_a is a subgroup of G and $|C(a)| = [G : Z_a]$.

Proof. If $x, y \in Z_a$, then $xya = xay = axy$ and $x^{-1}a = x^{-1}axx^{-1} = x^{-1}xax^{-1} = ax^{-1}$, so that $x, y, x^{-1} \in Z_a$. Hence, Z_a is a subgroup of G .

To prove the second part, note that each of the following statements is equivalent to the next:

- x and y belong to the same coset of Z_a in G ;
- $x^{-1}y \in Z_a$;
- $x^{-1}ya = ax^{-1}y$;
- $ya y^{-1} = xax^{-1}$.

In other words, the two elements x and y belong to the same coset of Z_a if and only if they conjugate a to the same element. Equivalently, the elements x and y belong to distinct cosets of Z_a in G if and only if they conjugate a to distinct elements of $C(a)$. Either way, the cosets of Z_a have been matched up in a one-to-one fashion with the elements of $C(a)$, so that $|C(a)| = [G : Z_a]$. ■

Note that in the quaternion group $xax^{-1} = a$ or e according as $x \in \{1, a, d, e\}$ or not. Thus, $|C(a)| = 2 = [G : Z_a]$, since we saw above that $Z_a = \{1, a, d, e\}$.

The *center* $Z(G)$ of the group G consists of all the elements of G which commute with every element of G . In other words,

$$Z(G) = \bigcap_{a \in G} Z_a.$$

It is clear that $Z(G) = G$ if and only if G is a commutative group. The following theorem provides a very useful tool in the search for the classification of groups.

Theorem 11.6 Let G be a finite group. Then there exist elements $a_1, a_2, \dots, a_k \in Z(G)$ such that

$$o(G) = o(Z(G)) + \sum_{i=1}^k [G : Z_{a_i}] \quad (11.7)$$

where, for each i , $[G : Z_{a_i}] > 1$ and $[G : Z_{a_i}] o(Z(G))$ is a proper divisor of $o(G)$.

Proof. As observed just prior to Proposition 11.5, every element of $Z(G)$ constitutes a conjugacy class by itself. Let $C(a_1), \dots, C(a_k)$ be a list of the other distinct conjugacy classes of G . Since each element of G belongs in some conjugacy class, we have

$$o(G) = o(Z(G)) + \sum_{i=1}^k |C(a_i)|.$$

Equation 11.7 now follows from Proposition 11.5.

If $[G : Z_{a_i}] = 1$, then $G = Z_{a_i}$, implying that a_i commutes with all the elements of G and contradicting the fact that $a_i \in Z(G)$. Thus, $[G : Z_{a_i}] > 1$ for each $i = 1, 2, \dots, k$.

Finally, since each $a_i \in Z(G)$ it follows that $Z(G)$ is a proper subgroup of each Z_{a_i} , and hence $[G : Z_{a_i}] o(Z(G))$ is a proper divisor of $[G : Z_{a_i}] o(Z_{a_i}) = o(G)$. ■

If G is the Quaternion group, $Z(G) = \{1, d\}$ and we can use $a_1 = a$, $a_2 = b$, and $a_3 = e$, since $C(a) = \{a, e\}$, $C(b) = \{b, f\}$, and $C(c) = \{c, g\}$.

Equation 11.7 is called the *class equation* of G . It has a surprising number of implications for finite groups. We demonstrate this first by classifying all the groups of order p^2 up to isomorphism. Note that this corollary is a very strong generalization of Proposition 9.17.

Corollary 11.8 Let G be a group of order p^2 , where p is a prime. Then G is isomorphic to either $(\mathbb{Z}_{p^2}, +)$ or to $(\mathbb{Z}_p, +)^2$.

Proof. If G is cyclic, it is isomorphic to $(\mathbb{Z}_{p^2}, +)$. Hence we may assume that every nonidentity element of G has order p . We first show that in the class equation of G , $k = 0$, so that $Z(G) = G$ and hence G is in fact commutative. To see this assume that $k > 0$ and let a_1, \dots, a_k be as in Theorem 11.6. Since $o(G) = p^2$, it follows that $[G : Z_{a_i}] = p$ for each $i = 1, 2, \dots, k$. Hence $o(Z(G))$ must also be divisible by p . This, however, contradicts the fact that $[G : Z_{a_i}] o(Z(G))$ is a proper divisor of $o(G) = p^2$. Thus, G is commutative.

Now that G is known to be a commutative group in which each nonidentity element has order p , let $a, b \in G$ be such that $\langle a \rangle$ and $\langle b \rangle$ are distinct subgroups of G . Since they both have prime order p , it follows that $\langle a \rangle \cap \langle b \rangle = \{1_G\}$ and so, by Proposition 11.2,

$$G \cong \langle a \rangle \times \langle b \rangle \cong (\mathbb{Z}_p, +) \times (\mathbb{Z}_p, +) = (\mathbb{Z}_p, +)^2. \quad \blacksquare$$

We are already familiar with two noncyclic groups of order p^2 for each prime p . These are the groups $\mathbb{Z}_p[x, \leq 1]$ and $(\text{GF}(p, P(x)), +)$ where $P(x)$ is any irreducible quadratic over \mathbb{Z}_p . Since all the nonzero elements of these groups have order p , it follows that they are isomorphic to $(\mathbb{Z}_p, +)^2$. The next theorem, above and beyond its utility in the classification of finite groups, also provides a partial converse to Lagrange's Theorem (Theorem 9.8) about the orders of subgroups. The actual converse to Lagrange's Theorem is false (Exercises 9.4.28 to 9.4.38).

Theorem 11.9 (Cauchy) If the order of the finite group G is divisible by the prime p , then G contains an element of order p .

Proof. We proceed by induction on $o(G)$. The theorem is trivial for groups of order one. Let the prime p be fixed, let n be a positive integer divisible by p and assume the

theorem to be true for all groups of order less than n . Let G be a group of order n and assume that G has no element of order p . By Proposition 9.7 we may assume that the order of every element of G is relatively prime to p .

Let a_1, a_2, \dots, a_k be as in Theorem 11.6. Since Z_{a_i} is a subgroup of G , it contains no element of order p . Since $Z_{a_i} \neq G$, it follows that $o(Z_{a_i}) < n$, so that by the induction hypothesis p cannot be a divisor of $o(Z_{a_i})$. However, $o(G) = [G : Z_{a_i}] o(Z_{a_i})$, and hence p must be a divisor of $[G : Z_{a_i}]$ for each $i = 1, 2, \dots, k$. It now follows from the class equation (Equation 11.7) that p divides $o(Z(G))$.

Let z be any nonidentity element of $Z(G)$ (z exists because $o(Z(G)) > 1$) and let $H = \langle z \rangle$. Since $Z(G)$ is commutative, it follows that H is a normal subgroup of $Z(G)$. Since p does not divide $o(H)$, it must divide $o(Z(G)/H) < n$. By the induction hypothesis, $Z(G)/H$ has some element, say bH , $b \notin H$, of order p . This means that $(bH)^p = H$ and hence $b^p \in H$. However, since $o(b)$ and p are relatively prime, it follows that there exist integers A and B such that $A o(b) + B p = 1$, and so

$$b = b^1 = b^{A o(b) + B p} = (1_G)^A (b^p)^B \in H,$$

which is a contradiction. Thus G must contain an element of order p . ■

We now have sufficient information to classify all the groups of order $2p$ where p is prime.

Corollary 11.10 If $p > 2$ is a prime and G is a group of order $2p$, then G is isomorphic to either $(\mathbb{Z}_{2p}, +)$ or D_p .

Proof. By Theorem 11.9, G has elements a and b of orders p and 2 , respectively. Since $H = \langle a \rangle$ has half the elements of G , it follows from Proposition 10.3 that H is a normal subgroup of G . Consequently, $bab^{-1} \in bHb^{-1} = H = \langle a \rangle$, and hence $bab^{-1} = a^k$ for some $k \in \mathbb{Z}_p$. However,

$$bakb^{-1} = (bab^{-1})(bab^{-1}) \cdots (bab^{-1}) = a^k a^k \cdots a^k = a^{k^2},$$

and so

$$a = b^2 a b^{-2} = b(bab^{-1})b^{-1} = ba^k b^{-1} = a^{k^2}.$$

It follows that $k^2 \equiv 1 \pmod{p}$ or $k \equiv \pm 1 \pmod{p}$.

If $k \equiv 1 \pmod{p}$, then $bab^{-1} = a$ or $ba = ab$. Since $o(a) = p$ and $o(b) = 2$, it follows from Proposition 9.7 that $o(ab) = 2p$ so that $G \cong (\mathbb{Z}_{2p}, +)$.

Otherwise, $k \equiv -1 \pmod{p}$ so that $bab^{-1} = a^{-1}$, or $ba = a^{-1}b = a^{p-1}b$. In this case G is necessarily isomorphic to D_p . To see this observe that the $2p$ elements of G can be listed as $a^0b^0, \dots, a^{p-1}b^0, a^0b^1, \dots, a^{p-1}b^1$. On the other hand, in D_p let $\alpha = (1\ 2\ \dots\ p)$ and let $\beta = (2\ p)(3\ p-1)\cdots((p-1)/2\ (p+1)/2)$. Then $\beta\alpha\beta^{-1} = \alpha^{-1}$, or $\beta\alpha = \alpha^{-1}\beta = \alpha^{p-1}\beta$. It is now easily verified that the function $f(a^i b^j) = \alpha^i \beta^j$ defines an isomorphism of G and D_p (Exercise 11.2.12). ■

Exercises 11.2

- Let G be a commutative group of order $p_1 p_2 \cdots p_r$ where p_1, p_2, \dots, p_r are distinct primes. Prove that $G \cong (Z_{p_1}, +) \times (Z_{p_2}, +) \times \cdots \times (Z_{p_r}, +)$.
- Prove that a finite group has exactly one conjugacy class if and only if it is trivial.
- Prove that a finite group has exactly two conjugacy classes if and only if it has order 2.
- Prove that a finite group has exactly three conjugacy classes if and only if it has order 3.
- Let p and q be distinct primes. Prove that every noncommutative group of order pq has a trivial center.
- Let k and n be arbitrary positive integers and let p be a prime such that $(k, p) = 1$ and $n/2 < p \leq n$. Let f be a function of n variables that have k distinct variants. Prove that there is a p -cycle σ in S_n such that $\sigma f = f$.
- Suppose G is a commutative group of order $p_1^{r_1} p_2^{r_2} \cdots p_h^{r_h}$ where p_1, p_2, \dots, p_h are distinct primes. Prove that G has subgroups G_1, G_2, \dots, G_h that have orders of $p_1^{r_1}, p_2^{r_2}, \dots, p_h^{r_h}$, respectively, such that $G \cong G_1 \times G_2 \times \cdots \times G_h$.
- Classify the following groups according to isomorphism type:

(a) $(\mathbb{Z}_6, +)$	(e) $\sqrt[6]{1}$
(b) $\langle (1\ 2\ 3\ 4\ 5\ 6) \rangle$	(f) $S_{3,f}$ where $f = x_1 x_2 x_3$
(c) $\langle (1\ 2\ 3)(4\ 5) \rangle$	(g) (\mathbb{Z}_9^*, \cdot)
(d) S_3	

9. Classify the following groups according to isomorphism type:

- | | |
|--|--|
| (a) $(\mathbb{Z}_8, +)$ | (f) $(\text{GF}(2, x^3 + x^2 + 1), +)$ |
| (b) D_4 | (g) $(\text{GF}^*(3, x^2 + x + 2), \cdot)$ |
| (c) the Quaternion group | (h) $(\mathbb{Z}_2, +) \times (\mathbb{Z}_4, +)$ |
| (d) $\sqrt[8]{1}$ | (i) $(\mathbb{Z}_2, +)^3$ |
| (e) $\langle (1\ 2\ 3\ 4\ 5\ 6\ 7\ 8) \rangle$ | |

10. Classify the following groups according to isomorphism type:

- | | |
|--|---|
| (a) $(\mathbb{Z}_9, +)$ | (c) $(\text{GF}(3, x^2 + x + 2), +)$ |
| (b) $\sqrt[7]{1}$ | (d) $\langle (1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9) \rangle$ |
| (e) $S_{9,f}$ where $f = x_1^2 x_2 + x_2^2 x_3 + \cdots + x_9^2 x_1$ | |
| (f) $(\mathbb{Z}_3, +)^2$ | |

11. Classify the following groups according to isomorphism type:

- | | | |
|---|---|--------------------|
| (a) $(\mathbb{Z}_{10}, +)$ | (b) D_5 | (c) $\sqrt[10]{1}$ |
| (d) $S_{5,f}$ where $f = x_1 x_2 + x_2 x_3 + x_3 x_4 + x_4 x_5 + x_5 x_1$ | | |
| (e) $(\mathbb{Z}_{11}^*, \cdot)$ | (f) $\langle (1\ 2\ 3\ 4\ 5)(6\ 7) \rangle$ | |

12. Prove that the function f defined in the proof of Corollary 11.10 is indeed an isomorphism.

13. Find a group of order 16 that is isomorphic to neither $(\mathbb{Z}_{16}, +)$ nor $(\mathbb{Z}_4, +)^2$.

Chapter Summary

The direct product of groups was introduced as a means for constructing a host of new groups. Cauchy's partial converse to Lagrange's Theorem and the class equation proved to be useful tools in investigating the isomorphism types of groups. It was shown that for any prime number p there are only two nonisomorphic groups of order p^2 and only two nonisomorphic groups of order $2p$. The consecutive solutions of certain landmark quadratic Diophantine equations leads mathematicians to question the traditional view of prime numbers as integers that cannot be factored into smaller numbers.

Chapter Review Exercises

Mark the following true or false.

1. The direct product of $(\mathbb{Z}_8, +)$ with S_4 has 32 elements.
2. $G \times \{1_G\} \cong G$.
3. If G is a group, then the conjugacy class $C(a)$ of any element a of G is a subgroup of G .
4. If G is a group and $a \in G$, then Z_a is a commutative group.
5. If G is a commutative group, then $Z_a = G$ for all $a \in G$.
6. Every group of order 49 is commutative.
7. If H is a normal subgroup of index 100 in G , then G/H contains an element of order 25.
8. If p is a prime number, then every group of order $2p$ is commutative.

New Terms

center, 266	conjugate, 265
centralizer, 265	decomposable group, 264
class equation, 267	direct product, 261
conjugacy class, 265	

Supplementary Exercises

1. Prove that every finite commutative group is the direct sum of cyclic groups.
2. Characterize all the finite groups that have exactly n conjugacy classes for as many positive integers n as possible.

Chapter 12



NUMBER THEORY

WE NOW EMBARK on a different journey, in pursuit of the ubiquitous rings and ideals. As was the case before (in Chapter 1) the journey began in ancient Egypt and Mesopotamia where someone discovered that some numbers are qualitatively different from others. The road then took us on to Greece, France, Switzerland, and Germany.

12.1 Pythagorean Triples

The Diophantine–Pythagorean equation $x^2 + y^2 = z^2$ generalizes in many ways as well as inspires many fruitful questions and very clever answers. The quadratic residue theorem is encountered along the way, and we explain why Gauss considered it to be the jewel in the crown of number theory. The proof given here makes strong use of the geometrical property of plane lattices.

The Theorem of Pythagoras, which states that in a right triangle with hypotenuse of length z , and shorter sides of lengths x and y , the equation

$$x^2 + y^2 = z^2 \tag{12.1}$$

holds, is considered to be one the most important theorems of mathematics. Examples of such solution triples are $\{3, 4, 5\}$, $\{5, 12, 13\}$, and $\{12,709, 13,500, 18,541\}$. The fact that the third of these was one of 15 such triples listed in the Babylonian tablet PLIMPTON 322, which dates between 1900 and 1650 BCE, attests to the fascination with which these triples were regarded, even as far back as 4,000 years ago. It was natural for the more mathematically minded scribes of the time to try to generate a list of all such triples. Euclid included part of this problem as Lemma I to Proposition 29 of Book X of *The Elements*:

Proposition 12.2 If $u > v$ are two positive integers, then $x = 2uv$, $y = u^2 - v^2$, and $z = u^2 + v^2$ satisfy $x^2 + y^2 = z^2$.

Proof. See Exercise 12.1.6. ■

The reason this solution was qualified above as only “partial” is that it failed to prove that all the desired triples can be so obtained. The statement and proof of this “converse” require some care. A *Pythagorean triple* is a list of positive integers (x, y, z) , which satisfy Equation 12.1 with the proviso that (x, y, z) and (y, x, z) are considered to be the same Pythagorean triples. A Pythagorean triangle is a triangle the length of whose sides constitute a Pythagorean triple. It is easy to see that the common factor of any two members of a Pythagorean triple must also divide the third one and hence the condition that x , y , and z share no common factor but 1 is tantamount to saying that every two of them are relatively prime. Such triples are said to be *primitive*. It is clear that if g is the greatest common divisor of the components x, y, z of a Pythagorean triple, then $(x/g, y/g, z/g)$ constitutes a primitive triple.

Lemma 12.3 If (x, y, z) is a primitive Pythagorean triple, then z is odd and exactly one of x and y is odd.

Proof. If x and y are even, it follows that z is also even, thus contradicting the primitivity of the triangle. Nor can both be odd, since otherwise there exist positive integers a , b , and c such that

$$(2a+1)^2 + (2b+1)^2 = (2c)^2$$

or $1+1 \equiv 0 \pmod{4}$, which is impossible. ■

Lemma 12.4 Let a , b , and c be positive integers such that $a^2 = bc$ and $(b, c) = 1$. Then there exist integers u and v such that $b = u^2$ and $c = v^2$.

Proof. Let

$$a = p_1^{r_1} p_2^{r_2} \cdots p_m^{r_m}$$

be the prime factorization of a . Then

$$a^2 = p_1^{2r_1} p_2^{2r_2} \cdots p_m^{2r_m}.$$

Since $(b, c) = 1$, it follows that after a suitable permutation of the subscripts

$$b = p_1^{2r_1} p_2^{2r_2} \cdots p_k^{2r_k} \quad \text{and} \quad c = p_{k+1}^{2r_{k+1}} p_{k+2}^{2r_{k+2}} \cdots p_m^{2r_m}$$

where k is some integer between 1 and m . Hence we can set

$$u = p_1^{r_1} p_2^{r_2} \cdots p_k^{r_k} \quad \text{and} \quad c = p_{k+1}^{r_{k+1}} p_{k+2}^{r_{k+2}} \cdots p_m^{r_m}. \quad \blacksquare$$

Theorem 12.5 A Pythagorean triple (x, y, z) , in which x is even, is primitive if and only if it has the form

$$x = 2st, \quad y = s^2 - t^2, \quad \text{and} \quad z = s^2 + t^2 \quad (12.6)$$

where s and t are positive integers of opposite parity, $s > t$, and $(s, t) = 1$. Every Pythagorean triple is an integer multiple of some primitive triple.

Proof. Let (x, y, z) be a primitive triple in which x is even. It follows that both y and z are odd so that $u = (z - y)/2$ and $v = (z + y)/2$ are positive integers such that

$$\left(\frac{x}{2}\right)^2 = \frac{z^2 - y^2}{4} = \frac{z - y}{2} \cdot \frac{z + y}{2} = uv.$$

Note that any common divisor of u and v is necessarily also a common divisor of y and z (because $y = v - u$ and $z = u + v$). Since $(y, z) = 1$, it follows that $(u, v) = 1$.

Now, the product of u and v is a perfect square $(x/2)^2$. Since u and v are relatively prime, they must each be perfect squares too, say $u = t^2$ and $v = s^2$. Thus

$$\left(\frac{x}{2}\right)^2 = uv \quad \text{or} \quad x^2 = 4s^2t^2 \quad \text{or} \quad x = 2st \quad (12.7)$$

and

$$y = v - u = s^2 - t^2 \quad \text{and} \quad z = u + v = s^2 + t^2. \quad (12.8)$$

The integers s and t must be of opposite parity since otherwise y and z would both be divisible by 2, contradicting the primitivity of the triple (x, y, z) . Finally $u < v$ and $(u, v) = 1$ imply that $s > t$ and $(s, t) = 1$.

Conversely, let s and t be positive integers of opposite parity such that $s > t$ and $(s, t) = 1$. It is easily verified that if x , y , and z are defined by means of Equations 12.7 and 12.8, then they form a Pythagorean triple with an even x . It remains to show that this triple is primitive. If p is any odd prime that is a common factor of both y and z , then p is also a common factor of the relatively prime

$$s^2 = \frac{z + y}{2} \quad \text{and} \quad t^2 = \frac{z - y}{2},$$

contradicting the fact that $(s, t) = 1$. As for the case $p = 2$, since s and t have different parities y and z are both odd, and so 2 is not a common factor of y and z either. Thus, y and z have no common prime factors, i.e., $(y, z) = 1$. ■

For example, if $s = 1$ and $t = 2$, then $x = 4$, $y = 3$, and $z = 5$. On the other hand, $s = 125$ and $t = 54$ yield $x = 13,500$, $y = 12,709$, and $z = 18,541$.

To find all Pythagorean triangles with a side of length 18, begin by finding all primitive Pythagorean triples one of whose legs is a divisor of 18. Because the parameters s and t in Theorem 12.5 are distinct, it follows that the length of every leg of any right triangle is at least

$$s^2 - t^2 \geq 4 - 1 = 3$$

and equals 3 if and only if $s = 2$ and $t = 1$, no right triangles have legs of length 1 or 2, and 3 is only in the Pythagorean triple $(3, 4, 5)$. This yields the triple $(18, 24, 30)$. Because 6 is even, it must be the side $2st$ where s and t have opposite parities. This, by inspection, is impossible. A similar argument eliminates 18. This leaves only 9 which must equal $s^2 - t^2$, since it clearly cannot equal either $2st$ or $s^2 + t^2$. The factorization

$$9 = s^2 - t^2 = (s + t)(s - t)$$

has only the solutions $s = 5$ and $t = 4$. This yields the triangle $2(9, 40, 41) = (18, 80, 82)$.

In order to find all the Pythagorean triangles (x, y, z) such that $40 \leq z \leq 50$, some tedious work leads to the conclusion that the only positive integers z which are the sums of two squares $s^2 + t^2 \leq 50$ where $(s, t) = 1$ and $s > t$ are

$5 = 2^2 + 1^2,$	$10 = 3^2 + 1^2,$
$13 = 3^2 + 2^2,$	$17 = 4^2 + 1^2,$
$25 = 4^2 + 3^2,$	$26 = 5^2 + 1^2,$
$29 = 5^2 + 2^2,$	$34 = 5^2 + 3^2,$
$37 = 6^2 + 1^2,$	$41 = 5^2 + 4^2.$

Of these, the only primitive triple in the range $40 \leq z \leq 50$ is $(9, 40, 41)$. To find nonprimitive solutions we locate for each integer d all triples $(\bar{x}, \bar{y}, \bar{z})$ such that $40 \leq d\bar{z} \leq 50$.

For $d = 2$ the range of \bar{z} becomes $20 \leq \bar{z} \leq 25$ and for this range only 25 is the sum of two squares, $25 = 4^2 + 3^2$. This gives the triple $2(7, 24, 25) = (14, 48, 50)$.

For $d = 3$ the range becomes $14 \leq \bar{z} \leq 16$ none of which numbers is the sum of two squares.

For $d = 4$ the range becomes $10 \leq \bar{z} \leq 12$ and for this range only 10 is the sum of two squares, $10 = 3^2 + 1^2$. This gives the triple $4(6, 8, 10) = (24, 32, 40)$.

For $d = 5$ the range becomes $8 \leq \bar{z} \leq 10$ and for this range only 10 is the sum of two squares with relatively prime sides. This gives the triple $5(6, 8, 10) = (30, 40, 50)$.

For $d = 6$ the range becomes $7 \leq \bar{z} \leq 8$, which yields no triples.

For $d = 7$ the range becomes $6 \leq \bar{z} \leq 7$, which yields no triples.

For $d = 8$ the range becomes $5 \leq \bar{z} \leq 6$, and for this range only 5 is the sum of two squares. This gives the triple $8(3, 4, 5) = (24, 32, 40)$.

For $d = 9$ the range narrows down to just $5 = 2^2 + 1^2$, which yields the triple $9(3, 4, 5) = (27, 36, 45)$.

For $d = 10$ the range becomes $4 \leq \bar{z} \leq 5$ and for this range only $5 = 2^2 + 1^2$ is the sum of two squares. This yields the triple $10(3, 4, 5) = (30, 40, 50)$.

The multiplier d cannot be greater than 10 because otherwise both the endpoints of the range are less than 5 which is the smallest positive number that can be expressed as the sum of two relatively prime and distinct squares. Hence the answer consists of the following list:

$$(30, 40, 50), (14, 48, 50), (27, 36, 45), (9, 40, 41), (24, 32, 40).$$

In conclusion, we note that Fermat took it for granted that Proposition 12.2 did completely account for all the Pythagorean triples. Theorem 12.5 was formulated and proved by Euler about 100 years later.

Exercises 12.1

1. Let (x, y, z) be a Pythagorean triple. Prove the following statements:

- (a) At least one of x and y is divisible by 3;
- (b) At least one of x , y and z is divisible by 5.

2. Find all Pythagorean triples (x, y, z) with
 - (a) $40 \leq \min\{x, y\} \leq 45$
 - (b) $45 \leq \max\{x, y\} \leq 50$
 - (c) $50 \leq z \leq 60$.
3. Find all the Pythagorean triangles having one side of length 481.
4. Find all Pythagorean triples (x, y, z) with $x < y$ and $z = 21$.
5. Determine the right triangles with integral sides whose areas equal their perimeter. (Hint: Work in terms of s and t .)
6. Prove Proposition 12.2.

12.2 Sums of Two Squares

In his book *Arithmetica*, written in the third century BCE, Diophantus included a variety of propositions regarding sums of squares, including Problem 8 of Book 2 which asks for the division of a square into two other squares. This problem is clearly inspired by the notion of Pythagorean triples. Fermat, coming upon this problem while reading the *Arithmetica* wrote in the book's margin a note to the effect that a cube cannot be split into two cubes nor a fourth power into two fourth powers. He claimed to have a remarkable proof of the following statement:

For any integer $n > 2$ there exist no positive integers x, y, z such that $x^n + y^n = z^n$,

the proof of which, alas, was too long for the margin. Fermat's proof, if indeed he had one, accompanied him to the grave. 350 years later a proof, hundreds of pages long, which made use of state-of-the-art mathematical tools and also relied on much mathematics developed in the intervening centuries, was given by Andrew Wiles and Richard Taylor.

One approach to obtaining Pythagorean triples could be to ask which positive integers are expressible as the sums of two squares? This problem was formulated by Fermat and solved, 100 years later, by Euler. We note in passing that every square is trivially also the sum of two squares, namely, itself and the square of "side" 0. The following proposition is, of course, very helpful in this context.

Proposition 12.9 (Brahmagupta 598–660) If each of two positive integers is the sum of two squares, then so is their product.

prime	sum of squares	prime	sum of squares
2	$1^2 + 1^2$	43	—
3	—	47	—
5	$1^2 + 2^2$	53	$2^2 + 7^2$
7	—	59	—
11	—	61	$5^2 + 6^2$
13	$2^2 + 3^2$	67	—
17	$1^2 + 4^2$	71	—
19	—	73	$3^2 + 8^2$
23	—	79	—
29	$2^2 + 5^2$	83	—
31	—	89	$5^2 + 8^2$
37	$1^2 + 6^2$	97	$4^2 + 9^2$
41	$4^2 + 5^2$		

Table 12.1 Sums of squares

Proof. Suppose $m = a^2 + b^2$ and $n = c^2 + d^2$. Then

$$\begin{aligned}
 mn &= (a^2 + b^2)(c^2 + d^2) = a^2c^2 + a^2d^2 + b^2c^2 + b^2d^2 \\
 &= (ac)^2 + (bd)^2 - 2(ac)(bd) + (ad)^2 + (bc)^2 + 2(ad)(bc) \\
 &= (ac - bd)^2 + (ad + bc)^2. \quad \blacksquare
 \end{aligned}$$

In view of Proposition 12.9 and the Fundamental Theorem of Arithmetic (Theorem 4.9), it should suffice to answer a restricted question: Which prime numbers are expressible as the sum of two squares? Some experimentation (see Table 12.1) would lead any interested student to correctly conjecture the following proposition.

Theorem 12.10 An odd prime p can be expressed as the sum of two squares if and only if there is a positive integer n such that $p = 4n + 1$.

The proof is somewhat intricate and is broken up into several lemmas.

Lemma 12.11 An odd prime of the form $4k + 3$ is not expressible as the sum of two squares.

Proof. Let $p = 4k + 3$ and suppose, by way of contradiction, that p is the sum of the two squares x^2 and y^2 . Since these two squares add up to the odd number $p = 4k + 3$, it follows that one of them, say x , is even and the other, that is, y , is odd. Hence there exist integers a and b such that

$$4n + 3 = (2a)^2 + (2b + 1)^2 = 4(a^2 + b^2 + b) + 1 \equiv 1 \pmod{4},$$

which is impossible since $4n + 3 \equiv 3 \pmod{4}$. ■

Lemma 12.12 If $(a, m) = 1$, then the equivalence $ax \equiv b \pmod{m}$ has a unique solution for each value of b .

Proof. This follows immediately from Corollary 4.4. ■

Fermat's Theorem (Theorem 5.15) states that if p is a prime and $a \not\equiv 0 \pmod{p}$, then

$$a^{p-1} \equiv 1 \pmod{p}.$$

It follows that

$$\left[a^{(p-1)/2} \right]^2 = a^{p-1} \equiv 1 \pmod{p}.$$

In other words $a^{(p-1)/2}$ is a solution of the equation $x^2 \equiv 1 \pmod{p}$. Since \mathbb{Z}_p is a field, this equation has only ± 1 as its solutions, and it follows that $a^{(p-1)/2} = \pm 1$.

Theorem 12.13 (Wilson's Theorem) If p is a prime, then $(p-1)! \equiv -1 \pmod{p}$.

Proof. Because \mathbb{Z}_p is a field, 1 and $p-1$ are the only solutions of the equivalence $x^2 \equiv 1 \pmod{p}$. In other words, 1 and -1 are the only elements of \mathbb{Z}_p that are their own inverses. Consequently

$$(p-1)! = 1 \cdot (p-1) \cdot (2 \cdot 2^{-1}) \cdot (3 \cdot 3^{-1}) \cdots (p-1)/2 \cdot [(p-1)/2]^{-1} \equiv -1 \pmod{p}. \quad \blacksquare$$

For example, $(3-1)! = 2 \equiv -1 \pmod{3}$, $(5-1)! = 24 \equiv -1 \pmod{5}$, and $(7-1)! = 720 \equiv -1 \pmod{7}$.

Modular arithmetic, especially over a field with a prime number of elements, has analogs of many notions of real arithmetic. It would be nice to have a notion of positive and negative, and indeed a fruitful definition of modular “positive” numbers can be

defined. Such is the concept of quadratic residue: The integer $a \not\equiv 0$ is said to be a *quadratic residue* modulo p if there is a solution in \mathbb{Z}_p to the equation

$$x^2 \equiv a \pmod{p}.$$

If no such solution exists, then a is a *quadratic nonresidue*.

Proposition 12.14 (Euler's Criterion) Let a be a positive integer and p be a prime such that $p \nmid a$. Then a is a quadratic residue modulo p if and only if

$$a^{(p-1)/2} \equiv 1 \pmod{p}. \quad (12.15)$$

Proof. The easy case where $p = 2$ is relegated to Exercise 12.2.1. Hence we may assume that p is odd. The proof adds a slight twist to the proof of Wilson's theorem: instead of pairing elements of \mathbb{Z}_p whose product is 1 we pair elements whose product is a . If a is a quadratic nonresidue, then the equivalence

$$m^2 \equiv a \pmod{p} \quad (12.16)$$

has no solutions, so that

$$-1 \equiv (p-1)! \equiv (1 \cdot a) \cdot (2 \cdot a2^{-1}) \cdots ((p-1)/2 \cdot a[(p-1)/2]^{-1}) \equiv a^{(p-1)/2}. \quad (12.17)$$

On the other hand, if a is a quadratic residue modulo p , then there exists an element $m \in \mathbb{Z}_p$ such that both m and $-m \equiv p-m \pmod{p}$ are distinct solutions of Equation 12.16. These roots of Equation 12.17 contribute a singleton each to the product in Equation 12.16 and hence in this case the product of Equation 12.17 simplifies to

$$-1 \equiv \cdots \equiv a^{(p-1-2)/2} \cdot m \cdot (p-m),$$

or

$$a^{(p-3)/2}(-a) \equiv -a^{(p-1)/2} \pmod{p},$$

and the desired equivalence (Equation 12.15) follows immediately. ■

For example, let $p = 7$. Then Equation 12.15 has a solution for $a = 1, 2, 4$ and does not for $a = 3, 5, 6$. At the same time, $(p-1)! = 6! \equiv -1 \pmod{7}$, $(7-1)/2 = 3$, and, modulo 7, $1^3 \equiv 2^3 \equiv 4^3 \equiv 1$, whereas $3^3 \equiv 5^3 \equiv 6^3 \equiv -1$.

Those values of a for which Equation 12.16 has a solution are called the quadratic residues of p . It is clear from the definition that

$$1^2, 2^2, 3^2, \dots, ((p-1)/2)^2 \pmod{p}$$

are all quadratic residues of p .

It follows from the above propositions that if p is prime, then the following three statements are equivalent:

- a is a quadratic residue mod p ;
- $x^2 \equiv a$ has a solution mod p ;
- $a^{(p-1)/2} \equiv 1 \pmod{p}$.

We add another equivalent statement:

Lemma 12.18 Let p be an odd prime. Then the congruence

$$x^2 + 1 \equiv 0 \pmod{p} \tag{12.19}$$

has a solution in \mathbb{Z}_p if and only if

$$p \equiv 1 \pmod{4}. \tag{12.20}$$

Proof. Suppose Congruence 12.19 has a solution in \mathbb{Z}_p . Then so, clearly, does

$$x^2 \equiv -1 \pmod{p},$$

which means that -1 is a quadratic residue modulo p . By Euler's criterion,

$$(-1)^{(p-1)/2} \equiv 1 \pmod{p}$$

and hence the exponent must be even. Thus, there is an integer k such that $(p-1)/2 = 2k$ or $p = 4k + 1$, which is tantamount to Congruence 12.24. The proof of the converse is similar and is relegated to Exercise 12.2.5. ■

Lemma 12.21 Let $p = 4k + 3$ be a prime and let np be a sum of two squares, say $np = a^2 + b^2$ for some $a, b \in \mathbb{Z}$, then

- p divides both a and b .
- p appears in the prime factorization of n with an even exponent.

Proof. The first statement is clear if p divides a . If p does not divide a , then a has a multiplicative inverse a^{-1} modulo p and it also follows from the hypothesis that

$$a^2 + b^2 \equiv 0 \pmod{p}.$$

Division by a^2 gives $(a^{-1}b)^2 + 1 \equiv 0 \pmod{p}$. In other words, $x = a^{-1}b$ is a solution of Equation 12.19 and hence $p \equiv 1 \pmod{4}$. This, however, contradicts the hypothesis $p = 4k + 3$.

For the second statement, suppose that p does not appear in the prime factorization of n with an even exponent. Fix p and let n be the least positive integer such that the exponent of p in the prime factorization of $np = a^2 + b^2$ is odd. By the first part of this proof, p divides both a and b ; and, of course, appears with positive even exponents in the prime factorizations of a^2 , b^2 , and n . Hence the terms of the equation

$$\frac{n}{p^2} = \left(\frac{a}{p}\right)^2 + \left(\frac{b}{p}\right)^2$$

are in fact integers rather than just rational numbers and so they constitute a smaller counterexample than the previous one. This contradicts the minimality of the first counterexample. This contradiction completes the proof of the lemma. ■

The proof of the second part of the above lemma is a variant of mathematical induction called the *method of infinite descent*. It was formulated by Fermat in the context of this very topic.

If x is any real number, then the *floor function* $\lfloor x \rfloor$ denotes the unique integer $\lfloor x \rfloor$ such that $\lfloor x \rfloor \leq x < \lfloor x \rfloor + 1$. Thus, $\lfloor 2.5 \rfloor = 2$, $\lfloor -2.5 \rfloor = -3$, and $\lfloor -0.5 \rfloor = -1$.

Theorem 12.22 Every prime of the form $4k + 1$ can be expressed as the sum of two relatively prime squares.

Proof. Let $p = 4k + 1$ be a prime. It follows from Euler's criterion (Proposition 12.14) that there is a number s such that $s^2 \equiv -1 \pmod{p}$. Consider the function of the integers x and y defined by $f(x, y) = sx - y$ for $0 \leq x, y < \sqrt{p}$. Since there are $\lfloor \sqrt{p} \rfloor + 1$ choices for each of x and y , the number of points in the domain of f is

$$(\lfloor \sqrt{p} \rfloor + 1)^2 > (\sqrt{p})^2 = p,$$

which is greater than the number of integers in the codomain of f . Consequently there are two distinct points (x_1, y_1) and (x_2, y_2) such that $f(x_1, y_1) = f(x_2, y_2)$, or $sx_1 - y_1 = sx_2 - y_2$, and so

$$sx \equiv y \pmod{p} \quad (12.23)$$

where $x = x_1 - x_2$ and $y = y_1 - y_2$. Note that if $x = 0$, then $y = 0$ and vice versa; if $y = 0$, then, since $s \not\equiv 0$, it follows that $x = 0$. Hence neither x nor y are zero. Squaring Congruence 12.23 yields $y^2 \equiv s^2 x^2 \equiv -x^2 \pmod{p}$, or $x^2 + y^2 \equiv 0 \pmod{p}$. It follows that $p \mid (x^2 + y^2)$. However,

$$0 < x^2 + y^2 < 2(\sqrt{p})^2 = 2p,$$

and so we conclude that

$$x^2 + y^2 = p. \quad (12.24)$$

The reason x and y are relatively prime is that every prime factor of both must also divide p and must therefore equal p . This, however, is impossible since it entails

$$x^2 + y^2 \geq p^2 + p^2 = 2p^2 > p,$$

thus contradicting Equation 12.24. ■

Proposition 12.25 The positive integer n is a sum of two squares if and only if every prime divisor of n of form $4k + 3$ appears in the unique factorization of n with an even exponent.

Proof. Suppose

$$n = 2^t p_1^{r_1} p_2^{r_2} \cdots p_d^{r_d} q_1^{s_1} q_2^{s_2} \cdots q_e^{s_e}$$

is expressible as the sum of two squares, where $p_i \equiv 1 \pmod{4}$ for $i = 1, 2, \dots, d$, and $q_j \equiv 3 \pmod{4}$ for $j = 1, 2, \dots, e$. By Lemma 12.21 each of the s_j 's is even.

Conversely, suppose each of the s_j 's is even. If $s_j = 2$, then we have the trivial expression $s_j^2 = s_j^2 + 0^2$. The application of Proposition 12.9 and Lemma 12.11 shows that the number n defined above is also presentable as the sum of two squares. ■

Exercises 12.2

1. Prove that Proposition 12.14 holds when $p = 2$.

$p = 4k + 1$	$p = 4k + 3$
$p = \square + \square$	$p \neq \square + \square$
$-1 \equiv a^2 \pmod{p}$	$-1 \not\equiv a^2 \pmod{p}$

Table 12.2 A summary of the section

- Let q be a prime of form $4k + 3$. Prove that q^2 is not the sum of two nonzero squares.
- Find a representation of 5,525 as the sum of two squares.
- Determine which of the following integers is a sum of two squares: 98, 343, 735, 1,428, and 4,680. Find such a representation if one exists.
- Complete the proof of Lemma 12.18.
- Prove that the prime p is of the form $4k + 1$ if and only if p divides $n^2 + (n + 1)^2$.
- Let $n = 2 \cdot 3^2 \cdot 5^3 \cdot 13^5$. Is there a Pythagorean triangle with hypotenuse n ?

12.3 Quadratic Reciprocity

The theorem that goes by the name of the Law of Quadratic Reciprocity has the distinction of having been reproved more times (over 200) than any other in mathematics, except for the Theorem of Pythagoras. Its validity was conjectured by both Euler in the decade 1742–1751 and by Adrien-Marie Legendre in 1785. It was proved by Gauss in 1801. Gauss considered it to be his *theorema aureum* (golden theorem) and over his lifetime he produced at least six different proofs for it. It has also been called a crown jewel of number theory. The proof presented here is based on one found on Wikipedia which, in turn is based on a proof of Eisenstein's.

Given an (odd) prime p and an arbitrary integer a , their *Legendre symbol* is defined as

$$\left(\frac{a}{p}\right) = \begin{cases} 1 & \text{if } a \text{ is a quadratic residue of } p; \\ -1 & \text{if } a \text{ is not a quadratic residue of } p; \\ 0 & \text{if } p \text{ divides } a. \end{cases}$$

For example, by the calculations immediately following the proof of Euler's criterion (Proposition 12.14),

$$\left(\frac{1}{7}\right) = \left(\frac{2}{7}\right) = \left(\frac{4}{7}\right) = 1$$

and

$$\left(\frac{3}{7}\right) = \left(\frac{5}{7}\right) = \left(\frac{6}{7}\right) = -1.$$

We begin with a list of easily proved properties of the Legendre symbol.

Theorem 12.26 Let p be an odd prime. Then

$$\left(\frac{a}{p}\right) \equiv a^{(p-1)/2} \pmod{p};$$

$$\left(\frac{a}{p}\right)\left(\frac{b}{p}\right) = \left(\frac{ab}{p}\right);$$

$$a \equiv b \pmod{p} \text{ implies that } \left(\frac{a}{p}\right) = \left(\frac{b}{p}\right);$$

$$\text{if } (a, p) = 1, \text{ then } \left(\frac{a^2 b}{p}\right) = \left(\frac{b}{p}\right);$$

$$\left(\frac{1}{p}\right) = 1;$$

$$\left(\frac{-1}{p}\right) = (-1)^{(p-1)/2}.$$

Proof. The first is a restatement of Euler's criterion. The proof of the second is relegated to Exercise 12.3.12. The third statement is clear, and the fourth follows the second. The last two also follow from Euler's criterion. ■

Theorem 12.27 (Fermat, Euler) If p is an odd prime, then

$$\left(\frac{2}{p}\right) = (-1)^{(p^2-1)/8}.$$

Proof. Let p be an odd prime. If $p = 4k + 1$ for some integer k , then, modulo p ,

$$\begin{aligned}
 2^{2k}(2k)! &= 2 \cdot 4 \cdot 6 \cdots 4k \\
 &= 2 \cdot 4 \cdot 6 \cdots 2k \cdot (2k+2) \cdot (2k+4) \cdots (4k-4) \cdot (4k-2) \cdot 4k \\
 &\equiv 2 \cdot 4 \cdot 6 \cdots 2k \cdot (-2k+1) \cdot (-2k+3) \cdots (-5) \cdot (-3) \cdot (-1) \\
 &= 2 \cdot 4 \cdot 6 \cdots 2k \cdot (-(2k-1)) \cdot (-(2k-3)) \cdots (-5) \cdot (-3) \cdot (-1) \\
 &= (-1)^k (2k)! \pmod{p}.
 \end{aligned}$$

Since $(2k)!$ is relatively prime to $p = 4k + 1$, it can be canceled out of the above congruence and we obtain

$$2^{(p-1)/2} = 2^{2k} \equiv (-1)^k \pmod{p}.$$

By Euler's criterion and the previous equation

$$\left(\frac{2}{p}\right) \equiv 2^{(p-1)/2} \equiv (-1)^{(p-1)/4} \pmod{p}.$$

On the other hand, if $p = 4k + 3$ then, modulo p ,

$$\begin{aligned}
 2^{2k+1}(2k+1)! &= 2 \cdot 4 \cdot 6 \cdots (4k+2) \\
 &= 2 \cdot 4 \cdot 6 \cdots (2k+2) \cdot (2k+4) \cdots (4k-2) \cdot 4k \cdot (4k+2) \\
 &\equiv 2 \cdot 4 \cdot 6 \cdots (2k+2) \cdot (-(2k+1)) \cdots (-5)(-3)(-1) \\
 &= (-1)^{k+1}(2k+1)! \pmod{p}.
 \end{aligned}$$

Since $(2k+1)!$ is relatively prime to $p = 4k + 3$, it can be canceled out of the above congruence and we obtain

$$2^{(p-1)/2} = 2^{2k+1} \equiv (-1)^{k+1} \pmod{p}.$$

By Euler's criterion,

$$\left(\frac{2}{p}\right) \equiv 2^{2k+1} \equiv (-1)^{(p+1)/4} \pmod{p}.$$

It remains to show that whenever $p = 4k + 1$,

$$\frac{p-1}{4} \equiv \frac{p^2-1}{8} \pmod{2},$$

and whenever $p = 4k + 3$,

$$\frac{p+1}{4} \equiv \frac{p^2-1}{8} \pmod{2}.$$

These tasks are relegated to Exercise 12.3.13. ■

The floor function is very useful in describing long division: If the number x is to be (long) divided by d , then

$$x = d\lfloor x/d \rfloor + r \tag{12.28}$$

where $r < d$ is the remainder.

Lemma 12.29 (Eisenstein) Let p and q be odd primes and let the variable u vary over the even numbers $E = \{2, 4, \dots, p-1\}$. Then

$$\left(\frac{q}{p}\right) = (-1)^{\sum_u \lfloor qu/p \rfloor}.$$

Proof. For each $u \in E$ let $r(u)$ denote the least positive residue of qu modulo p . For example, if $p = 17$ and $q = 13$, then u has the values 2, 4, 6, 8, 10, and 12 and $r(u)$ assumes the values 9, 1, 10, 2, 11, and 3.

We show that the function $(-1)^{r(u)}r(u)$ is in fact a permutation of E . The integers $(-1)^{r(u)}r(u)$ are all even (in our running example, they are 8, 16, 10, 2, 4, and 14). This is obvious when $r(u)$ is even. If $r(u)$ is odd, then $(-1)^{r(u)}r(u) = -r(u) < 0$. However, by definition, $0 \leq r(u) < p$ and hence the least positive residue of $r(u)$ modulo p is $r(u) + p$, which has even parity, contradicting the assumption that $r(u)$ is odd.

The integers in $r(u)$ are all distinct. For, if

$$(-1)^{r(u)}r(u) \equiv (-1)^{r(v)}r(v) \pmod{p},$$

then $u \equiv r(u) \equiv \pm r(v) \equiv \pm v \pmod{p}$. If $u \equiv -v$, then u and v are both even, p is odd, $3 \leq u + v \leq 2p - 3$ and $u + v \equiv 0 \pmod{p}$, which is only possible if $u + v = p$. Since u and v are even and p is an odd prime, we are forced to conclude that $u \equiv v \pmod{p}$. Since u and v are both in E , it follows that $u = v$.

The function $(-1)^{r(u)}r(u)$ is now known to have the set

$$\{2, 4, \dots, 2p-1\}$$

as its domain and its range, and is also known to be one-to-one. It therefore is a permutation (or rearrangement) of E . Recall that, by definition, $r(u) \equiv qu \pmod{p}$ and hence we have the following chain of equivalences modulo p :

$$\begin{aligned} 2 \cdot 4 \cdots (p-1) &\equiv (-1)^{r(2)}r(2) \cdot (-1)^{r(4)}r(4) \cdots (-1)^{r(p-1)}r(p-1) \\ &\equiv (-1)^{r(2)}2q \cdot (-1)^{r(4)}4q \cdots (-1)^{r(p-1)}(p-1)q \\ &\equiv (-1)^{r(2)+r(4)+\cdots+r(p-1)} \cdot 2 \cdot 4 \cdots (p-1)q^{(p-1)/2} \end{aligned}$$

and hence, upon division by $2 \cdot 4 \cdots (p-1)$, we have

$$\left(\frac{q}{p}\right) \equiv q^{(p-1)/2} \equiv (-1)^{r(2)+r(4)+\cdots+r(p-1)} \pmod{p} = (-1)^{\sum r(u)} \quad (12.30)$$

where u varies over E .

However, (long) division of qu by p yields

$$qu = p \left\lfloor \frac{qu}{p} \right\rfloor + r(u).$$

Sum this equation over all $u \in E$ to obtain

$$q \sum u = \sum p \left\lfloor \frac{qu}{p} \right\rfloor + \sum r(u).$$

Since all the u 's are even, it follows that

$$p \sum \left\lfloor \frac{qu}{p} \right\rfloor \equiv \sum r(u) \pmod{2}.$$

Since p is odd,

$$\sum \left\lfloor \frac{qu}{p} \right\rfloor \equiv \sum r(u) \pmod{2}.$$

In view of Equation 12.30, we are done. ■

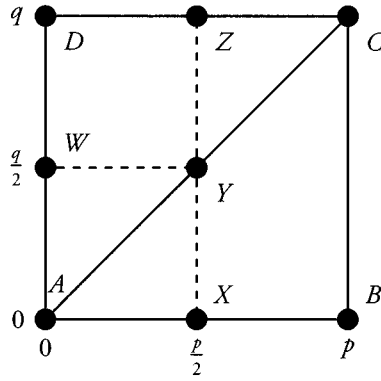


Figure 12.1 Lattice point diagram

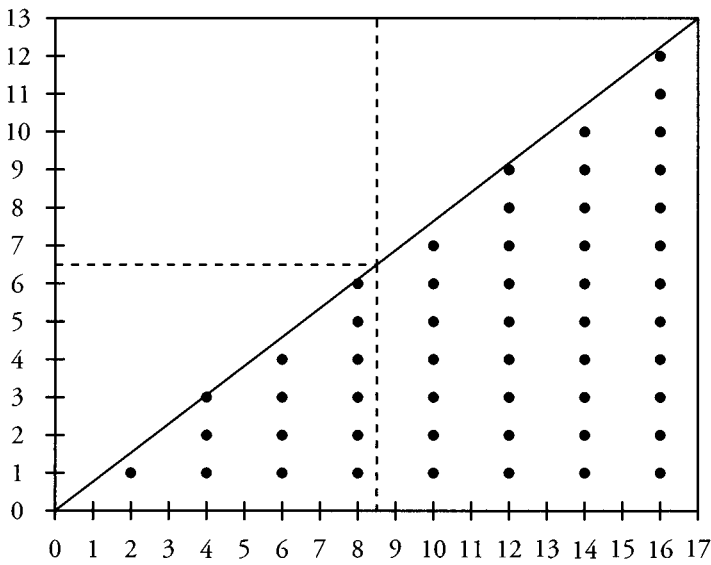


Figure 12.2 Example showing lattice points inside ABC with even x -coordinates for $p = 11$ and $q = 13$

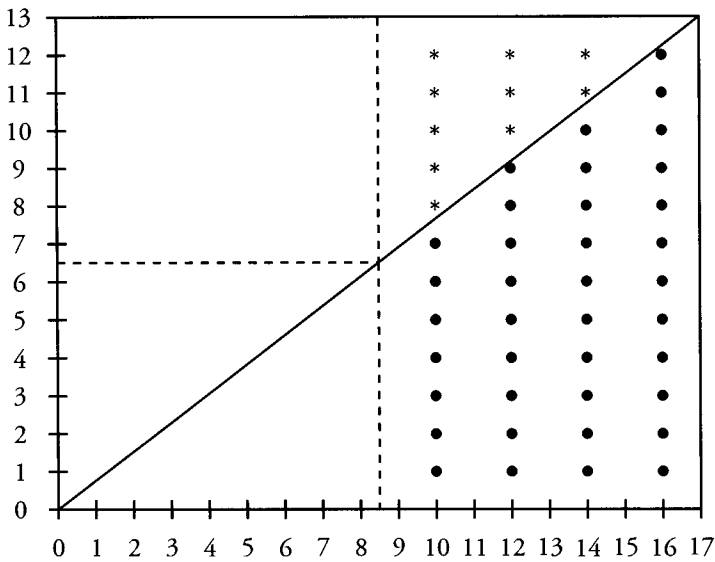


Figure 12.3 The number of points with even x -coordinate inside $BCYX$ is equal modulo 2 to the number of such points in CZY

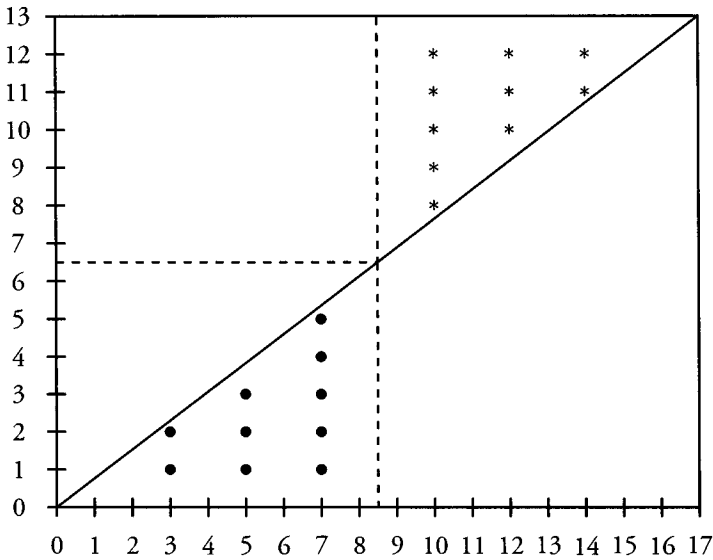


Figure 12.4 The number of points with even x -coordinate inside CZY is equal to the number of points with odd x -coordinate inside AXY

Yet another definition is called for here. A point (x, y) with integer components is said to be a *lattice point*. If S is any subset of the plane, then $E(S)$ and $O(S)$ denote the number of lattice points in S whose x -coordinate, respectively, is even or odd. Also, $L(S) = E(S) + O(S)$ is the total number of lattice points in E .

Proposition 12.31 (Eisenstein) If p and q are distinct odd primes then

$$\left(\frac{q}{p}\right) = (-1)^{\sum_{j=1}^{(p-1)/2} \lfloor \frac{jq}{p} \rfloor}.$$

Proof. Let p and q be two distinct odd primes and let

$$A = \{0, 0\}, \quad B = \{p, 0\}, \quad C = \{p, q\}, \quad D = \{0, q\}$$

be the rectangle of Figure 12.1 (where $p = 17$ and $q = 13$). It follows from Eisenstein's Lemma 12.29 that

$$\begin{aligned} \left(\frac{q}{p}\right) &= (-1)^{E(ABC)} = (-1)^{E(AXY)} \cdot (-1)^{E(BCYX)} \\ &\equiv (-1)^{E(AXY)+E(CZY)} \pmod{2} \end{aligned}$$

because

$$(-1)^{E(BCYX)+E(CZY)} = (q-1) \frac{p-1}{2} \equiv 0 \pmod{2}$$

and so, applying Eisenstein's Lemma 12.29 again, we obtain, modulo 2,

$$\left(\frac{q}{p}\right) \equiv (-1)^{E(AXY)+O(AXY)} = (-1)^{L(AXY)} = \sum_{j=1}^{(p-1)/2} \lfloor jq/p \rfloor \quad \blacksquare$$

Theorem 12.32 (Law of Quadratic Reciprocity) If p and q are distinct odd primes, then

$$\left(\frac{q}{p}\right) = (-1)^{[(p-1)/2][(q-1)/2]}.$$

Proof. Referring back to Figures 12.1 to 12.4,

$$\left(\frac{q}{p}\right) = (-1)^{L(AXY)}.$$

By symmetry,

$$\left(\frac{p}{q}\right) = (-1)^{L(AWY)}.$$

Since the diagonal AC contains no other lattice points than A and C , it follows that

$$\frac{p-1}{2} \frac{q-1}{2} = L(AWYX) = L(AXY) + L(AWY)$$

and hence, by Lemma 12.31,

$$(-1)^{[(p-1)/2][(q-1)/2]} = (-1)^{L(AXY)}(-1)^{L(AWY)} = \left(\frac{q}{p}\right)\left(\frac{p}{q}\right) \quad \blacksquare$$

Exercises 12.3

1. Consider $ax^2 + bx + c \equiv 0 \pmod{p}$ where p is an odd prime and $p \nmid a$. Let $D = b^2 - 4ac$. Prove that the congruence has
 - (a) no solutions of D if a is a quadratic nonresidue,
 - (b) a unique solution if $p \mid D$,
 - (c) exactly two solutions if D is a quadratic residue of p .
2. Suppose a is a quadratic residue of the odd prime p , and prove that $-a$ is also a quadratic residue of p if and only if -1 is a quadratic residue of p .
3. What are the least positive residues of the quadratic residues of 31?
4. Is $x^2 \equiv -2 \pmod{263}$ solvable?
5. Let p be an odd prime. Prove that the product of the quadratic residues of p is congruent to -1 or 1 modulo p according as $p \equiv 1 \pmod{4}$ or $p \equiv 3 \pmod{4}$.
6. Suppose p is a prime of the form $8k + 3$. Does p divide $2^{(p-1)/2} - 1$?
7. Characterize the odd primes $p \neq 7$ such that $x^2 \equiv 7 \pmod{p}$ has a solution in p .
8. Characterize the odd primes $p \neq 11$ such that $x^2 \equiv 11 \pmod{p}$ has a solution in p .
9. Calculate the following by hand:

(a) $\left(\frac{70}{97}\right)$	(b) $\left(\frac{-14}{83}\right)$	(c) $\left(\frac{263}{331}\right)$	(d) $\left(\frac{461}{773}\right)$
----------------------------------	-----------------------------------	------------------------------------	------------------------------------

10. Calculate the following by hand:

$$(a) \left(\frac{170}{1109} \right) \quad (b) \left(\frac{-1108}{1933} \right) \quad (c) \left(\frac{1263}{3313} \right) \quad (d) \left(\frac{2461}{4783} \right)$$

11. Let $p > 3$ be an odd prime. Prove that the sum of the quadratic residues of p is divisible by p . (Hint: Recall that $\sum_{i=1}^n i^2 = n(n+1)(2n+1)/6$.)

12. Prove the second part of Theorem 12.26.

13. Complete the proof of Lemma 12.29.

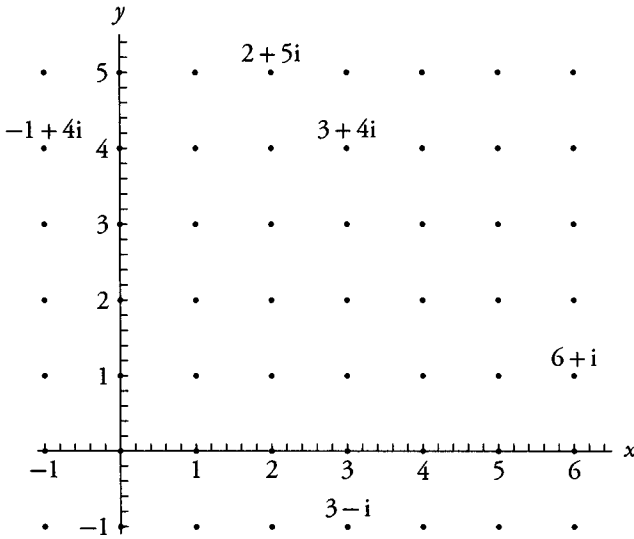
12.4 The Gaussian Integers

Having proved the Law of Quadratic Reciprocity, Gauss went on to look for higher-order analogs. To be specific, he worked on cubic and biquadratic (alias quartic) versions. Amongst other things he discovered that the results and their proofs were considerably simplified by widening the scope to complex “integers”, rather than restricting them to the standard integers of \mathbb{Z} . We now turn to the study of this extension.

The *Gaussian integers*, so named for obvious reasons, consist of all the complex numbers of the form $a + bi$ where $a, b \in \mathbb{Z}$ (see Figure 12.5). This new number system is denoted by $\mathbb{Z}[\sqrt{-1}]$, consistently with the definition of $F[x]$ in Section 6.1, and $17 + 9i$, $12 - 14i$, $-14i$, and 5 are all Gaussian integers. It was only natural for Gauss to seek out the primes of this new number system. Even the elementary fact $5 = (2 + i)(2 - i)$ is sufficient to indicate that, as mathematicians like to say, *something* is going on here.

However, some preliminary remarks are in order before we delve into the issue of the factorization of the Gaussian integers. Note that we stand in some danger of confusion as 5 seems to be a prime in the context of \mathbb{Z} but not so in the wider context of $\mathbb{Z}[\sqrt{-1}]$. To clear this up we henceforth designate the classical integers, \mathbb{Z} , as the *rational integers* and their primes as *rational primes*. In contrast the *Gaussian primes* are those of $\mathbb{Z}[\sqrt{-1}]$. The reason for the first appellation is that the primes of \mathbb{Z} do not involve $i = \sqrt{-1}$, which, by analogy with $\sqrt{2}$ and $\sqrt{3}$ are *irrational* numbers.

In this textbook we have so far discussed the issue of factorization in two contexts: the integers \mathbb{Z} and the polynomials $\mathbb{Q}[x]$. For the purpose of uniqueness of factorization of the integers it was necessary to regard $-p$ and p as redundant. Similarly, when factoring polynomials we consider $cP(x)$ and $P(x)$ to be essentially the same factors. These observations can be unified by defining the *units* of \mathbb{Z} to be 1 and -1 and the units of $\mathbb{Q}[x]$ to be all the nonzero rational numbers, \mathbb{Q}^* .

Figure 12.5 The Gaussian integers $\mathbb{Z}[\sqrt{-1}]$

Notice that the units of these two systems have the property that the inverse of a unit is again a unit, and we take this to be the defining characteristic of the units of $\mathbb{Z}[\sqrt{-1}]$. Specifically, the units of the Gaussian integers are the numbers 1, -1 , i , and $-i$. Two Gaussian integers z and w are said to be *associates* provided there is a unit u such that $uz = w$. For example, $2 - 3i$, $3 + 2i$, $-2 + 3i$, and $-3 - 2i$ are all associates of each other. In general associates come in groups of four, 0 being the only exception.

The Gaussian integer π is a Gaussian prime provided that whenever $\pi = zw$ is a factorization of π , then either z or w is a unit. We now define the ultra useful *norm* function $N \rightarrow \mathbb{Z}[i]$ via

$$N(a + bi) = a^2 + b^2 = |a + bi|^2.$$

The following lemma provides a very strong, though still incomplete, tool for identifying the Gaussian primes:

Lemma 12.33 The norm function is multiplicative in the sense that $N(zw) = N(z)N(w)$ for all $z, w \in \mathbb{C}$.

Proof. This follows from Proposition 12.9 and the fact that for any complex number z , $N(z) = |z|^2$. ■

Lemma 12.34 The units of $\mathbb{Z}[\sqrt{-1}]$ are characterized by the property that their norm is 1.

Proof. Suppose u is a unit. Then there exists another Gaussian integer, say w , such that $uw = 1$ and hence $N(uw) = N(u)N(w) = 1$ where $N(u)$ and $N(w)$ are positive rational integers. Consequently, $N(u) = 1$.

Conversely, let $N(u) = 1$. Hence, if $u = a + bi$, then $a^2 + b^2 = 1$ where a and b are rational integers. The solutions of this equation are $u = 1$, $u = i$, $u = -1$, and $u = -i$. Since each of the last four numbers has an inverse, it follows that u is indeed a unit. ■

Table 12.3 lists factorizations for all the Gaussian integers with norm at most 50 and which lie in the first quadrant. Prime Gaussian integers are most easily recognized by their norms:

Corollary 12.35 If z is a Gaussian integer whose norm is a rational prime, then z is a Gaussian prime.

Proof. Suppose $N(zw)$ is a rational prime p . Then $N(p) = N(zw) = N(z)N(w)$. It follows that either $N(z) = 1$ or $N(w) = 1$ which implies that either z or w is a unit. Hence z is a Gaussian prime. ■

For example, $N(1 - i) = 2$, $N(1 + 2i) = 5$, $N(-3 + 2i) = 13$, $N(4 + i) = 17$, and $N(5 - 2i) = 29$. Hence $1 - i$, $1 + 2i$, $-3 + 2i$, $4 + i$, and $5 - 2i$ are all Gaussian primes.

Proposition 12.36 Let p be an odd rational prime. Then p is also a Gaussian prime if and only if p has the form $4k + 3$, if and only if p does not have an expression as the sum of two squares.

Proof. Let p be a rational prime of form $4k + 3$. If p is not a Gaussian prime, then there exist nonunit integers α and β such that $p = \alpha\beta$. Taking norms, $p^2 = N(\alpha)N(\beta)$. Since α and β are nonunits, both must have norm p . This means that if $\alpha = a + bi$, then $p = a^2 + b^2$, contradicting Theorem 12.10.

Alternatively, if $p = 4k + 1$, then there exist rational integers a and b such that $p = a^2 + b^2$. So if we set $\pi = a + bi$, then $\pi\bar{\pi} = p$ and so we have a factorization of p into nonunits. In other words, p is not a Gaussian prime. ■

The readers are reminded that the Euclidean algorithm plays a central role in the factorization of the rational integers and we now set out to show that an analogous procedure exists for the Gaussian integers. This is far from obvious as it will soon be

norm	factors	norm	factors	norm	factors
2	$1+i$	18	$3+3i=3(1+i)$	36	$6=-3i(1+i)^2$
4	$2=-i(1+i)^2$	20	$2+4i=(1+i)^2(2-i)$	37	$1+6i$
	$5+2i$		$4+2i=-i(1+i)^2(2+i)$		$6+i$
5	$1+2i$	25	$3+4i=(2+i)^2$	40	$2+6i=-i(1+i)^3(2+i)$
	$2+i$		$4+3i=i(2-i)^2$		$6+2i=-i(1+i)^3(2-i)$
9	3		$5=(2+i)(2-i)$	41	$4+5i$
10	$1+3i=(1+i)(2+i)$	26	$1+5i=(1+i)(3+2i)$		$5+4i$
	$3+i=(1+i)(2-i)$		$5+i=(1+i)(3-2i)$	45	$3+6i=3i(2-i)$
13	$2+3i$	29	$2+5i$		$6+3i=3(2+i)$
	$3+2i$		$5+2i$	49	7
16	$4=-(1+i)^4$	32	$4+4i$	50	$1+7i=i(1+i)(2-i)^2$
17	$1+4i$	34	$3+5i=(1+i)(4+i)$		$5+5i=(1+i)(2+i)(2-i)$
	$4+i$		$5+3i=(1+i)(4-i)$		$7+i=-i(1+i)(2+i)^2$

Table 12.3 Prime factorizations in $\mathbb{Z}[i]$

demonstrated that there are many number systems very similar to the Gaussian integers which do not possess such an analog.

Let α , β , and γ be Gaussian integers such that $\alpha\beta = \gamma$. Then we say that α is a *divisor* of γ and γ is a *multiple* of α . The integer γ is a greatest common divisor (GCD) or highest common factor (HCF) of α and β provided that

- (a) γ divides both α and β , and
- (b) γ is a multiple of every common divisor of α and β .

If such an HCF indeed exists, then it is denoted by (α, β) .

Lemma 12.37 (Division Algorithm) Let z and w be two nonzero Gaussian integers such that $N(z) \geq N(w)$. Then there exist two Gaussian integers q and r such that

$$z = qw + r \quad \text{and} \quad N(r) < N(w). \quad (12.38)$$

Moreover,

$$(z, w) = (w, r) \quad (12.39)$$

and there exist Gaussian integers λ and μ such that

$$(z, w) = \lambda z + \mu w. \quad (12.40)$$

Proof. Let $z = a + bi$ and $w = c + di$. Then

$$\frac{z}{w} = \frac{a + bi}{c + di} = s + ti$$

where

$$s = \frac{ac + bd}{c^2 + d^2} \quad \text{and} \quad t = \frac{bc - ad}{c^2 + d^2}.$$

Let s' and t' be the rational integers closest to s and t , respectively. Then clearly

$$|s - s'| \leq \frac{1}{2}, \quad \text{and} \quad |t - t'| \leq \frac{1}{2}. \quad (12.41)$$

Thus, we could think of $q = s' + t'i$ as the Gaussian integer that is closest to z/w . Set

$$\rho = \frac{z}{w} - q \quad (12.42)$$

and note that

$$N(\rho) = (s - s')^2 + (t - t')^2 \leq \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 < 1.$$

By Equation 12.41 $z = qw + \rho w$ where z , w , and δ are all Gaussian integers, so that ρw is also a Gaussian integer for which

$$N(\rho w) = N(\rho)N(w) < 1 \cdot N(w) = N(w).$$

So if we set $r = \rho w$ we are done.

As for Equation 12.39, it follows from $z = qw + r$ that every common factor of z and w is also a common factor of w and r , and, vice versa, every common factor of w and r is also a common factor of z and w . The desired equality follows immediately. Finally, the existence of λ and μ follows from the same argument that was used in Proposition 4.1 to prove the existence of A and B . ■

Proposition 12.43 For any two Gaussian integers α and β , not both 0, the HCF (α, β) exists and every two HCFs of α and β are associates.

Proof. Let I be the set of all the Gaussian integers of the form $\varphi\alpha + \psi\beta$ where $\varphi, \psi \in \mathbb{Z}[\sqrt{-1}]$. Let δ be an element of I of minimum positive norm, and suppose $\delta = \lambda\alpha + \mu\beta$. We will show that δ is an HCF of α and β .

First, note that by the Division Algorithm there exist κ and ρ , with $N(\rho) < N(\delta)$ such that

$$\rho = \alpha - \kappa\delta = \alpha - \kappa(\lambda\alpha + \mu\beta) = (1 - \kappa\lambda)\alpha + (-\kappa\mu)\beta$$

with, moreover, $N(\rho) < N(\delta)$. This is impossible unless $\rho = 0$. This, in turn, implies that $\delta \mid \alpha$ and similarly $\delta \mid \beta$, and so δ is a common divisor of α and β (first property). Finally, if γ divides both α and β , it must divide every integer combination of α and β , including δ (second property). ■

This proposition is the Gaussian analog of the process of division in the context of the rational integers and it is customary to refer to q and r as the *quotient* and *remainder*, respectively, resulting from the division of q by r . Note that in the Gaussian case there are occasionally some arbitrary choices called for. For example, if (s, t) is the center of the unit square that contains it, then any of the four surrounding points will serve as (s', t') . The outcome is the same regardless of the choice (why?).

Let m and n be two Gaussian integers such that $N(m) \geq N(n)$. Set $m_1 = m$ and $n_1 = n$, and let q_1 and r_1 be the appropriate quotient and remainder so that $m_1 = q_1 n_1 + r_1$. For $i = 1, 2, 3, \dots$ we set $m_{i+1} = n_i$ and $n_{i+1} = r_i$ with q_{i+1} and r_{i+1} being the respective quotient and remainder when m_{i+1} is divided by n_{i+1} . Note that if $r_i \neq 0$, then either

$r_{i+1} = 0$ or else $N(r_{i+1}) < N(r_i) = N(r_i)$. Because all the $N(r_i)$'s are rational integers this process must eventually produce the first k for which $r_k = 0$. Thus $(m, n) = r_{k-1}$.

For example, let $z = 5 - i$ and $w = 4 + 2i$. Then

$$\frac{z}{w} = \frac{5-i}{4+2i} = \frac{9}{10} - \frac{7i}{10}.$$

Consequently $r' = 1$, $s' = -1$, and $q = 1 - i$, from which we obtain $\rho w = z - qw = 5 - i - (1 - i)(4 + 2i) = -1 + i$. In other words, $(5 - i, 4 + 2i) = -1 + i$.

For another example, we will find a GCD of $-15 - i$ and $-9 + 5i$. Successive applications of the division algorithm yield the following equations:

$$-15 - i = (1 + 2i)(-9 + 5i) + (4 + 12i),$$

$$-9 + 5i = i(4 + 12i) + (3 + i),$$

$$4 + 12i = (2 + 3i)(3 + i) + (1 + i),$$

$$3 + i = (2 - i)(1 + i).$$

Hence the required GCD is $1 + i$.

Proposition 12.44 If g is a greatest common divisor of the two Gaussian integers z and w , then there exist Gaussian integers A and B such that $Am + Bn = g$.

Proof. See proofs of Propositions 4.1 and 6.14. ■

Proposition 12.45 Let π be a Gaussian prime and z and w Gaussian integers such that $\pi \mid zw$. Then either $\pi \mid z$ or $\pi \mid w$.

Proof. Suppose π does not divide z . Since π is a Gaussian prime that does not divide z , it follows that $(\pi, z) = 1$. Let A and B be Gaussian integers such that $A\pi + Bz = 1$. Multiply by w to get $A\pi w + Bzw = w$. Since π divides each of the summands on the left it follows that π divides their sum w . ■

Theorem 12.46 Every nonzero and nonunit Gaussian integer can be factored into Gaussian primes in an essentially unique way.

Proof. The proof of the existence of a factorization into primes proceeds by mathematical induction on the norm. The integers $\pm 1 \pm i$ are all the Gaussian integers of norm 2. Since they are also primes, the induction has been anchored. Assume the existence of a prime factorization has been demonstrated for all numbers of norm less than n , and let

z be a number with $N(z) = n$. If z is composite, then there exist Gaussian integers z_1 and z_2 , both nonzero and nonunits such that $z = z_1 z_2$, $N(z) = N(z_1)N(z_2)$, and, hence, $N(z_1) < N(z) = n$ and $N(z_2) < N(z) = n$. By the induction hypothesis, both z_1 and z_2 have prime factorizations which, together, yield a prime factorization of z . By induction, the existence of a prime factorization has been demonstrated.

We next turn to the issue of uniqueness. The proof proceeds again by induction on the norm. If z has norm 2, then, by Lemma 12.33, z is a prime and so there cannot be another factorization into primes. This anchors the induction process. Assume that the uniqueness has been established for all integers of norm less than n and that $N(z) = n$. Let

$$p_1^{h_1} p_2^{h_2} \cdots p_r^{h_r} = z = q_1^{k_1} q_2^{k_2} \cdots q_s^{k_s}$$

be two essentially distinct prime factorizations of z . Since p_1 divides n , it follows from a repeated application of Proposition 12.45 that $p_1 = \eta q_i$ for some subscript i and some unit η . It follows that $z/p_1 = z/\eta q_i$. Relabeling, if necessary, rewrite $z/p_1 = z/q_1$. Since the common norm of both sides of this equation is $N(z) < N(q_1/p_1)$, it follows from the induction hypothesis that $r = s$ and for $i = 1, 2, \dots, r$, $h_i = k_i$, and p_i and q_i are associates. Hence, by induction, we are done. ■

Lemma 12.47 For each Gaussian prime π there exists a rational prime p such that $\pi \mid p$.

Proof. Let π be a Gaussian prime and set $n = N(\pi) = \pi\bar{\pi}$. Let $n = p_1 p_2 \cdots p_k$ be a factorization of n into rational primes. Clearly $\pi \mid n$. Hence, by an easy extension of Proposition 12.45, there exists a rational prime p_i such that $\pi \mid p_i$. ■

Theorem 12.48 (Gauss, 1801) The Gaussian integer π is a Gaussian prime if and only if one of the following holds:

- (a) π is $1 - i$ or an associate;
- (b) π is a rational prime of the form $4k + 3$ or an associate;
- (c) $N(\pi) = p$, where p is a rational prime of the form $4k + 1$.

Proof. Let π have one of the formats listed above. We show that in all three cases π is a Gaussian prime.

If π is $1 - i$ or an associate, then π is a Gaussian prime because its norm is 2 (see Corollary 12.35).

Next, let π be a rational prime of form $4k + 3$ and suppose $\pi = \alpha\beta$ is a factorization of π . Then $N(\pi) = \pi^2 = N(\alpha)N(\beta)$. Since π has the form $4k + 3$, it is not the sum of

two squares and hence $N(\alpha) \neq \pi \neq N(\beta)$, and it follows that either $N(\alpha) = 1$ or $N(\beta) = 1$. Hence one of α or β is a unit and so π is Gaussian prime.

For the third case, let p be the sum of two nonzero squares, say a^2 and b^2 so that $p = (a + bi)(a - bi) = N(a + bi) = N(a - bi)$ and set $\pi = a + bi$. Then π and its associates have prime norm and hence, by Corollary 12.35, they are Gaussian primes.

Conversely, we now argue that if $\pi = a + bi$ is any Gaussian prime, then it must fall into one of these three categories. If $b = 0$, then $\pi = a$ is a rational prime as well as a Gaussian prime and by Theorem 12.10 p has the form $4k + 3$. If $a = 0$, then $\pi = bi$ and so π is the associate of a rational prime of the form $4k + 3$. In both of these cases π falls into the second category.

Hence it may be assumed that neither a nor b vanishes. Set $p = a^2 + b^2$. We now show that p is necessarily a rational prime. For otherwise there exist nonunit rational integers c and d such that

$$(a + bi)(a - bi) = p = cd$$

which leads to two distinct (Gaussian) prime factorizations of p . Hence, if $N(\pi) = 2$, then π falls in the first category. Otherwise π has form $4k + 1$ and it falls in the third category. ■

Proposition 12.49 If p is a rational prime of the form $4k + 1$, then p is expressible as the sum of two squares in a unique way.

Proof. The existence of such an expression was established in Theorem 12.10. Suppose by way of contradiction, that there exist a rational prime p and positive rational integers s , t , x , and y such that $\{s, t\} \neq \{x, y\}$ and $s^2 + t^2 = p = x^2 + y^2$. Then, the classification (Theorem 12.48) implies that

$$(s + it)(s - it) = p = (x + iy)(x - iy)$$

are two distinct Gaussian prime factorizations of p , contradicting Lemma 12.11. Hence, each rational prime of form $4k + 1$ is expressible as the sum of two primes in only one way. ■

We now show how the preceding material can be used to solve some Diophantine equations.

Corollary 12.50 The equation $x^2 + 1 = y^3$ has the unique integer solution $x = 0$, $y = 1$.

Proof. Let x, y be an integer solution of the proposed equation. Let δ be any common, nonunit, Gaussian divisor of $x + i$ and $x - i$. Then δ must divide $(x + i) - (x - i) = 2i$. A glance at the factorizations of Table 12.3 tells us that δ is an associate of $1 + i$. Hence

$$(1 + i) \mid (x \pm i) \quad (12.51)$$

By Lemma 12.33, $N(1 + i) \mid N(x \pm i)$, or $2 \mid (x^2 + 1)$. Hence $x^2 + 1$ is even so that x must be odd.

However, if x is odd, we have $x^2 + 1 \equiv 2 \pmod{4}$ and so $y^3 \equiv 2 \pmod{4}$. This is impossible because the odd parity of x also implies that y would be even, in which case

$$y^3 \equiv 2^3 \left(\frac{y}{2}\right)^3 \equiv 0^3 \left(\frac{y}{2}\right)^3 \equiv 0 \pmod{4},$$

thus contradicting the previous equation.

Consequently, δ is a unit and hence $x + i$ and $x - i$ are relatively prime Gaussian integers or, in other words, they share no prime factors. Since their product is a cube (Gaussian integer), it follows that each of the two is a cube so that

$$x + i = (u + vi)^3 = u^3 - 3uv^2 + (3u^2v - v^3)i$$

where u and v are real integers. Separating real and imaginary parts yields the simultaneous equations $x = u^3 - 3uv^2$ and $1 = v(3u^2 - v^2)$. The second equation implies that $v = -1$ from which it follows that $u = 0$. Substituting into the first equation yields $x = 0$ and hence $y = 1$. ■

We note that the unique factorization property of $\mathbb{Z}[\sqrt{-1}]$ was used twice in the above proof, first to conclude that $x + i$ and $x - i$ are relatively prime and second to conclude that they are cubes of Gaussian integers.

Exercises 12.4

1. Find the number of all the Gaussian integers with norm less than n for $n = 5, 6, 7, 8, 9$
2. Let $g(n)$ denote the number of Gaussian integers of norm less than n . Is the number of integers n for which $g(n) = g(n + 1)$ finite or infinite?
3. Report on Gauss's circle problem.

4. Solve the Diophantine equation $x^2 + 1 = y^4$.
5. Solve the Diophantine equation $x^2 + 1 = y^5$.
6. Factor 100 into Gaussian primes.
7. Factor $37 - 9i$ into Gaussian primes.
8. Factor $62 + 41i$ into Gaussian primes.
9. Factor $1010 + 620i$ into Gaussian primes.
10. Factor $537 + 266i$ into Gaussian primes.
11. Apply the division algorithm of Lemma 12.37 to $z = 8 - 2i$ and $w = 3 + i$.
12. Apply the division algorithm of Lemma 12.37 to $z = 25 - 5i$ and $w = 2 + i$.
13. Use the Euclidean algorithm to find the GCD of the following pairs of Gaussian integers:

(a) $(3 + 11i, 4 - 2i)$	(c) $(33 + 19i, -16 + 37i)$
(b) $(14 + 29i, 19 + 26i)$	(d) $(26 + 19i, 29 + 14i)$

12.5 Eulerian Integers and Others

It stands to reason that the same technique that was used in the resolution of the preceding proposition could also be brought to bear on such a Diophantine equation as

$$y^3 = x^2 + 2. \quad (12.52)$$

In such a solution the right-hand side could be factored as

$$x^2 + 2 = (x + \sqrt{-2})(x - \sqrt{-2}).$$

This requires a context for such expressions as $x \pm \sqrt{-2}$, one which is easily provided by defining the *Eulerian integers* as

$$\mathbb{Z}[\sqrt{-2}] = \{ u + v\sqrt{-2} \mid u, v \in \mathbb{Z} \}.$$

It is easily verified (Exercise 12.5.1) that this set is closed with respect to addition, subtraction, and multiplication. See Table 12.4 for a list of small primes of $\mathbb{Z}[\sqrt{-2}]$.

Let us assume for the moment that $\mathbb{Z}[\sqrt{-2}]$ has unique prime factorization and see whether or not the uniqueness of such a factorization would lead to the solution. Here

$$x^2 + 2 = (x + \sqrt{-2})(x - \sqrt{-2})$$

and we must show that the two factors on the right are relatively prime in $\mathbb{Z}[\sqrt{-2}]$.

Suppose, by way of contradiction that δ is a prime in $\mathbb{Z}[\sqrt{-2}]$ that divides both factors. Then $\delta \mid 2\sqrt{-2}$. Since $2\sqrt{-2} = -\sqrt{-2}^3$ is a prime power, δ must be $\pm\sqrt{-2}$. Consequently $\sqrt{-2} \mid y^3$ and hence $\sqrt{-2} \mid y$. Taking norms we conclude that $2 \mid y$. This, however, implies that $x^2 \equiv 2 \pmod{4}$, which is impossible.

Finally let us solve the equation

$$x + \sqrt{-2} = (u + v\sqrt{-2})^3 = u^3 + 3u^2v\sqrt{-2} + 3uv^2(-2) + v^3(-2)\sqrt{-2}.$$

Separation of real and imaginary parts yields $x = u(u^2 - 6v^2)$ and $1 = v(3u^2 - 2v^2)$. The second equation implies that $v = \pm 1$ and $u = \pm 1$. This gives $x = 5$ and $y = \pm 3$ as the only solutions of Equation 12.52.

Of course, it is still necessary to prove that $\mathbb{Z}[\sqrt{-2}]$ does possess the unique prime factorization property, and we hasten to do so. Observe that the elements of $\mathbb{Z}[\sqrt{-2}]$ form a rectangular lattice in which each cell is a rectangle of dimensions $1 \times \sqrt{2}$ (see Figure 12.6).

Lemma 12.53 The restriction to $\mathbb{Z}[\sqrt{-2}]$ of the function

$$N(a + b\sqrt{-2}) = a^2 + 2b^2 \quad \text{for all } a, b \in \mathbb{C}$$

is multiplicative.

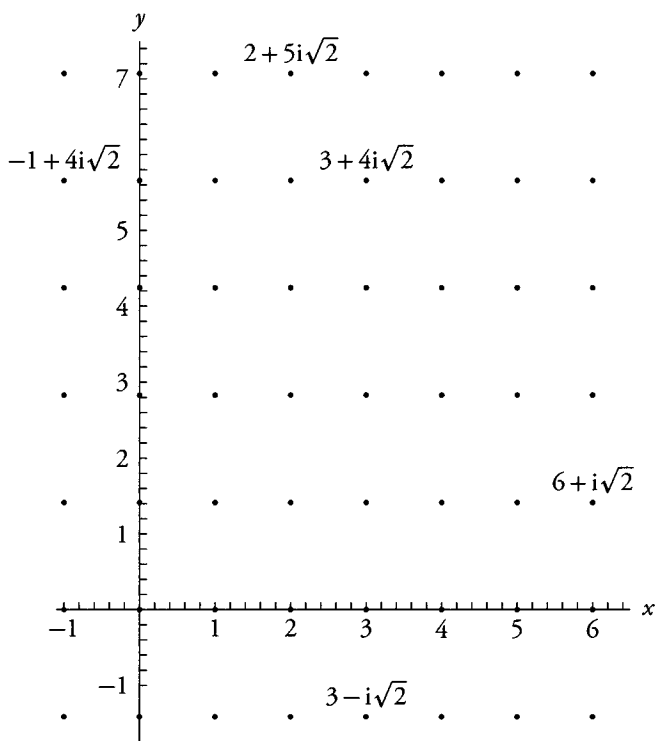
Proof. See Exercise 12.5.17. ■

The proof that $\mathbb{Z}[\sqrt{-2}]$ has a division algorithm is very similar to that of the Eulerian integers.

Lemma 12.54 (Division Algorithm) Let z and w be two nonzero Eulerian integers such that $N(z) \geq N(w)$. Then there exist two Eulerian integers q and r such that

$$z = qw + r \quad \text{and} \quad N(r) < N(w). \quad (12.55)$$

Moreover, $(z, w) = (w, r)$.

Figure 12.6 The Eulerian integers $\mathbb{Z}[\sqrt{-2}]$

Proof. Let $z = a + b\sqrt{-2}$ and $w = c + d\sqrt{-2}$. Then

$$\frac{z}{w} = \frac{a + b\sqrt{-2}}{c + d\sqrt{-2}} = s + t\sqrt{-2}$$

where

$$s = \frac{ac + 2bd}{c^2 + 2d^2} \quad \text{and} \quad t = \frac{bc - ad}{c^2 + 2d^2}.$$

Let s' and t' be rational integers that are closest to s and t , respectively. Then clearly

$$|s - s'| \leq \frac{1}{2}, \quad |t - t'| \leq \frac{\sqrt{2}}{2}$$

Thus, we could think of $q = s' + t'\sqrt{-2}$ as the Eulerian integer that is closest to z/w .

Set

$$\rho = \frac{z}{w} - q \tag{12.56}$$

and note that

$$N(\rho) = (s - s')^2 + (t - t')^2 \leq \left(\frac{1}{2}\right)^2 + \left(\frac{\sqrt{2}}{2}\right)^2 < 1.$$

By Equation 12.55 $z = qw + \rho w$ where z , w , and q are all Eulerian integers, so that ρq is also a Eulerian integer for which

$$N(\rho w) = N(\rho)N(w) < 1 \cdot N(w) = N(w).$$

So if we set $r = \rho w$, we are done with Equation 12.55.

As for this theorem's last equation, it follows from $z = qw + r$ that every common factor of z and w is also a common factor of w and r , and, vice versa, every common factor of w and r is also a common factor of z and w . The desired equality follows immediately. ■

For example, to apply the Euclidean division algorithm to $z = -15 - \sqrt{-2}$ and $w = 2 - 3\sqrt{-2}$, here

$$s + t\sqrt{-2} = (-24/22) - (47/22)i$$

and hence

$$q = s' + t'\sqrt{-2} = -1 - 2\sqrt{-2}$$

and

$$r = z - qw = (-15 - \sqrt{-2}) + (1 + 2\sqrt{-2})(2 - 3\sqrt{-2}) = -1.$$

For another example, to find the GCD of $(-115 - 31\sqrt{-2})$ and $(54 - 27\sqrt{-2})$, the first iteration of the Euclidean algorithm is

$$-115 - 31\sqrt{-2} = (-1 - \sqrt{-2})(54 - 27\sqrt{-2}) + (-7 - 4i)$$

and the second one

$$54 - 27\sqrt{-2} = (-2 + 5\sqrt{-2})(-7 - 4i).$$

Hence the required GCD is $-7 - 4i$.

Proposition 12.57 $\mathbb{Z}[\sqrt{-2}]$ has unique factorization.

It is known that for every odd prime p of the form $8k + 1$ or $8k + 3$ the equation $p = x^2 + 2y^2$ has a solution. It follows from the above proposition that this solution is unique.

The mathematical structure $\mathbb{Z}[\sqrt{-3}]$ is defined in a manner similar to the Gaussian integers $\mathbb{Z}[\sqrt{-1}]$ and the Eulerian integers $\mathbb{Z}[\sqrt{-2}]$. It is tempting to speculate that this structure also enjoys unique factorization. However, the argument of Lemma 12.37 fails in this new context. In fact, the equation $\mathbb{Z}[\sqrt{-3}]$ does not have unique factorization as is made clear by Table 12.5 and the factorizations

$$4 = 2 \times 2 = (1 + \sqrt{-3})(1 - \sqrt{-3}).$$

Euler, as well as other mathematicians of the time, was unaware of this fact and, as a result, produced several faulty proofs, including one that purported to show that the equation $x^3 + y^3 = z^3$ has no nonzero integer solutions. The structure $\mathbb{Z}[\sqrt{-3}]$ can be visualized just as $\mathbb{Z}[\sqrt{-2}]$ was, with the sole difference being that the tiling rectangle has height $\sqrt{3}$ rather than $\sqrt{2}$.

Exercises 12.5

1. Find the number of all the Eulerian integers with norm less than 5, 6, 7, 8 or 9.
2. Let $f(n)$ denote the number of Eulerian numbers of norm less than n . Is the number of values n for which $f(n) = f(n+1)$ finite or infinite?
3. Report on an Eulerian version of Gauss's circle problem.
4. Solve the Diophantine equation $x^2 + 2 = y^4$.
5. Solve the Diophantine equation $x^2 + 2 = y^5$.
6. Factor 100 into Eulerian primes.
7. Factor $-8 + 7\sqrt{-2}$ into Eulerian primes.
8. Factor $-22 - \sqrt{-2}$ into Eulerian primes.
9. Factor $-26 + 43\sqrt{-2}$ into Eulerian primes.
10. Factor $-394 - 526\sqrt{-2}$ into Eulerian primes.
11. Apply the division algorithm of Lemma 12.54 to $z = 5 + 3\sqrt{-2}$ and $w = 2 + \sqrt{-2}$.
12. Apply the division algorithm of Lemma 12.54 to $z = 25 - 5\sqrt{-2}$ and $w = 2 + \sqrt{-2}$.

norm	factors	norm	factors	norm	factors
2	$\sqrt{-2}$	18	$4 + \sqrt{-2} = \sqrt{-2}(1 + \sqrt{-2})^2$	34	$4 + 3\sqrt{-2} = \sqrt{-2}(3 - 2\sqrt{-2})$
3	$1 + \sqrt{-2}$		$3\sqrt{-2} = (1 + \sqrt{-2}) \cdot (1 - \sqrt{-2}) \cdot \sqrt{-2}$	36	$6 = \sqrt{-2}^4 (1 + \sqrt{-2})^2 \cdot (1 - \sqrt{-2})^2$
4	$2 = -\sqrt{-2}^2$	19	$1 + 3\sqrt{-2}$		$2 + 4\sqrt{-2} = \sqrt{-2}^2 (1 - \sqrt{-2})^2$
6	$2 + \sqrt{-2} = \sqrt{-2}(1 - \sqrt{-2})$	24	$4 + 2\sqrt{-2} = -\sqrt{-2}^3 (1 - \sqrt{-2})$	38	$6 + \sqrt{-2} = \sqrt{-2}(1 + 3\sqrt{-2})$
8	$-(\sqrt{-2})^3$	25	5	41	$3 + 4\sqrt{-2}$
9	$1 + 2\sqrt{-2} = -(1 - \sqrt{-2})^2$	27	$3 + 3\sqrt{-2} = (1 + \sqrt{-2})^2 \cdot (1 - \sqrt{-2})$	43	$5 + 3\sqrt{-2}$
11	$3 + \sqrt{-2}$		$5 + \sqrt{-2} = -(1 - \sqrt{-2})^3$	44	$6 + 2\sqrt{-2} = -\sqrt{-2}^2 (3 + \sqrt{-2})$
12	$2 + 2\sqrt{-2} = -\sqrt{-2}^2 (1 + \sqrt{-2})$	32	$4 = \sqrt{-2}^4$	48	$4 + 4\sqrt{-2} = \sqrt{-2}^4 (1 + \sqrt{-2})$
16	$4 = \sqrt{-2}^4$	33	$1 + 4\sqrt{-2} = (1 + \sqrt{-2}) \cdot (3 + \sqrt{-2})$	49	7
17	$3 + 2\sqrt{-2}$		$5 + 2\sqrt{-2} = (1 + \sqrt{-2}) \cdot (3 - \sqrt{-2})$	50	$5\sqrt{-2}$

Table 12.4 Some prime factorizations in $\mathbb{Z}[\sqrt{-2}]$

norm	prime	norm	prime
3	$\sqrt{-3}$	25	5
4	2	31	$2 + 3\sqrt{-3}$
4	$1 + \sqrt{-3}$	37	$5 + 2\sqrt{-3}$
7	$2 + \sqrt{-3}$	43	$4 + 3\sqrt{-3}$
13	$1 + 2\sqrt{-3}$	61	$7 + 2\sqrt{-3}$
19	$4 + \sqrt{-3}$		

Table 12.5 Some prime numbers of $\mathbb{Z}[\sqrt{-3}]$

13. Use the Euclidean algorithm to find the GCD of the following pairs of Eulerian integers:
- (a) $(11\sqrt{-2}, 5 - 2\sqrt{-2})$
 - (b) $(8 + 29\sqrt{-2}, 14 + 26\sqrt{-2})$
 - (c) $(45 + 19\sqrt{-2}, -44 + 37\sqrt{-2})$
 - (d) $(32 + 19\sqrt{-2}, 34 + 14\sqrt{-2})$
14. Let $g(n)$ denote the number of Eulerian primes of norm less than n . Is the number of values n for which $f(n) = f(n+1)$ finite or infinite?
15. Since $\mathbb{Z}[\sqrt{-3}]$ does not have unique factorization, the procedure used in the proof of Lemma 12.37, applied to this number system, must fail to produce a Euclidean algorithm. Explain where the failure occurs.
16. Prove that if

$$\omega = \frac{-1 + \sqrt{-3}}{2}$$

then $\mathbb{Z}[\omega]$ has unique factorization. (Hint: show that $\omega^2 = -1 - \omega$.)

17. Prove Lemma 12.53.

12.6 What Is the Essence of Primality?

In his groundbreaking article of 1847 *On the Theory of Complex Numbers*, E. E. Kummer (1810–1893) wrote:

I have observed, however, that even though $f(\alpha)$ cannot in any way be broken up into complex factors, it still does not possess the true nature of a complex prime number, for, quite

norm	divisors	norm	divisors
4	2	49	$2 + 3\sqrt{-5}$
5	$\sqrt{-5}$	7	
6	$1 + \sqrt{-5}$	54	$3 + 3\sqrt{-5} = 3(1 + \sqrt{-5})$
9	$2 + \sqrt{-5}$		$7 + \sqrt{-5} = (2 - \sqrt{-5}) \cdot (1 + \sqrt{-5})$
	3	56	$6 + 2\sqrt{-5} = 2(3 + \sqrt{-5})$
14	$3 + \sqrt{-5}$	61	$4 + 3\sqrt{-5}$
16	$4 = 2^2$	64	$8 = 2^3$
20	$2\sqrt{-5}$	69	$7 + 2\sqrt{-5}$
21	$1 + 2\sqrt{-5}$		$8 + \sqrt{-5}$
	$4 + \sqrt{-5}$	70	$5 + 3\sqrt{-5} = \sqrt{-5}(3 - \sqrt{-5})$
24	$2 + 2\sqrt{-5} = 2(1 + \sqrt{-5})$	80	$4\sqrt{-5} = 2^2\sqrt{-5}$
25	$5 = -\sqrt{-5}^2$	81	$1 + 4\sqrt{-5} = -(2 - \sqrt{-5})^2$
29	$3 + 2\sqrt{-5}$	84	$2 + 4\sqrt{-5} = 2(1 + 2\sqrt{-5})^2$
30	$5 + \sqrt{-5}$		$8 + 2\sqrt{-5} = 2(4 + \sqrt{-5})^2$
36	$6 = (1 + \sqrt{-5})(1 - \sqrt{-5})$	86	$9 + \sqrt{-5}$
	$6 = 2 \cdot 3$	89	$3 + 4\sqrt{-5}$
41	$6 + \sqrt{-5}$	94	$7 + 3\sqrt{-5}$
45	$3\sqrt{-5}$	96	$4 + 4\sqrt{-5} = 2^2(1 + \sqrt{-5})$
	$5 + 2\sqrt{-5} = \sqrt{-5}(2 - \sqrt{-5})$	100	$10 = 2^2\sqrt{-5}^4$
46	$1 + 3\sqrt{-5}$		

Table 12.6 Some prime factorizations in $\mathbb{Z}[\sqrt{-5}]$

commonly, it lacks the first and most important property of prime numbers; namely, that the product of two prime numbers is divisible by no other prime numbers.

The readers have already encountered this “most important” property as Proposition 12.43, which can be summarized as

$$p \mid ab \Rightarrow p \mid a \text{ or } p \mid b, \quad (12.58)$$

where a and b are arbitrary (rational) integers. Most introductory books on elementary number theory begin with some definition of the integers, go on to prove the Euclidean

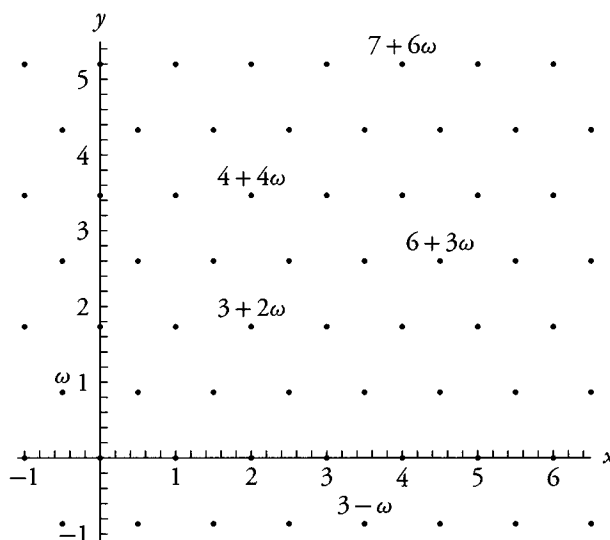


Figure 12.7 The Eisenstein Integers

algorithm, and then use this algorithm to prove Lemma 12.37. In Euclid's Book VII, Proposition 30 says

If two numbers by multiplying one another make some number and any prime measure¹ the product, it will also measure one of the original numbers.

This is clearly equivalent to Property 12.58. Some analog of Lemma 4.6 or Proposition 12.43 is then commonly used to deduce the Fundamental Theorem of Arithmetic (Theorem 4.9 and Proposition 6.9). It must be stressed here that for reasons unknown Euclid did not state the Fundamental Theorem of Number Theory explicitly. There is evidence to suggest that we today evaluate the subjective question of "which propositions are important and which are uninteresting" differently from Euclid.

There are, however, other number systems that admit of primes as well. As noted above, Gauss studied the number system bearing his name in depth, going so far as to give a detailed classification of its primes. Just as the Law of Biquadratic (or Quartic) Reciprocity produced the Gaussian integers as a by-product, so did the search for a Law

¹ Divide without remainder.

of Cubic Reciprocity motivate the definition of the system

$$\mathbb{Z}[\omega] = \{a + b\omega \mid a, b \in \mathbb{Z}\}$$

where ω is the cubic root of unity defined in Figure 12.7. In fact, Kummer greatly generalized these two number systems when he worked on the factorization of the elements of

$$\mathbb{Z}[\zeta] = \{a_0 + a_1\zeta + \cdots + a_{n-1}\zeta^{n-1} \mid a_i \in \mathbb{Z}\}$$

where ζ is any n -th root of unity. There he discovered a surprising fact: for $n = 24$ the number system $\mathbb{Z}[\zeta]$ does not have the desirable quality of unique factorization. The demonstration is rather technical and would take us far afield. Fortunately a much simpler example is available. In $\mathbb{Z}[\sqrt{-3}]$ we have the following distinct prime factorizations of 4:

$$2 \cdot 2 = (1 + \sqrt{-3})(1 - \sqrt{-3}).$$

This equation is disturbing because it implies that $2 \mid (1 + \sqrt{-3})(1 - \sqrt{-3})$ even though $2 \nmid 1 \pm \sqrt{-3}$ which runs counter to Property 12.58.

One way to describe what Kummer did is to say that when he realized that the prime numbers were “misbehaved” he replaced them with “new and improved” versions. To be precise, he changed the definition of primes to the following: A number p is prime if whenever it divides a product ab it must divide either a or b . That the rational primes of \mathbb{Z} have this property follows immediately from Lemma 4.6. The converse is also true and its proof is relegated to Exercise 12.6.1.

For the remainder of this book we shall refer to the traditional primes that have nondivisibility as their defining characteristic as *indecomposable integers*. For example in $\mathbb{Z}[\sqrt{-3}]$, where

$$2 \cdot 2 = (1 - \sqrt{-3})(1 + \sqrt{-3}),$$

each of the factors is an indecomposable nonprime number. They are indecomposable because of Table 12.5 (where they are listed as primes). To see that they are no longer to be considered as primes note that each has norm 4, but $\mathbb{Z}[\sqrt{-3}]$ has no elements of norm 2.

However, unique factorization is a very desirable quality in number systems and Kummer’s solution was to enlarge the context by positing the existence of a multitude of ideal numbers in each such system. Technically speaking, these ideal numbers were undefined,

only their interactions with the original integers in the system in question are specified. More precisely, Kummer defined what it means to say that a number is *divisible* by an ideal prime number.² This was fairly standard practice in the nineteenth-century mathematical world and examples of this type of reasoning are common: such was the case in this text with the complex numbers and the Galois fields. Kummer, Galois, and Euler were not concerned with the “real nature” of their creations, just with how to use them in mathematical reasoning.

The second half of the nineteenth century witnessed a shift toward absolute rigor amongst mathematicians. As part of this trend Dedekind suggested that in some contexts, at least, new elements could be represented by sets of old ones. The reader may be familiar with the method of Dedekind cuts developed by Dedekind which defines irrational real numbers with certain sets of rational numbers. A similar tack was taken by Dedekind for the interpretation of Kummer’s ideal numbers. They should be sets of rational integers. These are the *ideals* of the next chapter.

In the event, the mathematical establishment declined Kummer’s approach, though not his mathematical results, in favor of Dedekind’s approach which is the content of the next chapter. These doubts about the legitimacy of ideal integers notwithstanding, Kummer’s work is justly lauded for its originality, depth, and influence. His “definition” of ideal numbers will be discussed in greater detail in the next chapter.

Exercises 12.6

1. Let $q > 1$ be a positive integer of \mathbb{Z} such that for any two integers a and b , $q \mid ab$ implies that either $q \mid a$ or $q \mid b$. Show that q is a prime of \mathbb{Z} .
2. State and prove the converse of Exercise 12.6.1. Let $n > 1$ and $k \geq 1$ be positive integers with the following property: For every $a, b \in \mathbb{Z}$, if $n \mid a^k b^k$ then either $n \mid a^k$ or $n \mid b^k$. Then there exists a prime integer p such that $n \in \{p, p^2, \dots, p^k\}$.
3. Prove that the positive integer $n > 1$ is divisible by a square if and only if there exists a number m such that $n \nmid m$ and $n \mid m^2$.
4. State and prove an analog of Exercise 12.6.3 for divisibility by p^k .

Chapter Summary

The topics of Pythagorean triples, Sums of Squares, Quadratic Reciprocity, Gaussian Integers, and Eulerian Integers are each natural generalizations (or variations) on each

²See Edwards, p. 325.

other. All have surprising ties to geometry. They also motivate a reexamination of the notions of primeness and irreducibility.

Chapter Review Exercises

Mark the following true or false.

1. There is a positive integer n that belongs to infinitely many Pythagorean triples.
2. There is a prime number p that is expressible as the sum of two squares in infinitely many ways.
3. If p is a rational prime, so is $2p$.
4. $\left(\frac{-243}{333}\right) = 1$.
5. A killion is a number so large that it will kill you.

New Terms

associates, 295	norm, 295
Eulerian integers, 304	primitive, 274
Gaussian integers, 294	Pythagorean triple, 274
Gaussian primes, 294	quadratic nonresidue, 281
indecomposable integers, 313	quadratic residue, 281
irrational, 294	rational integers, 294
lattice point, 292	rational primes, 294
Legendre symbol, 285	units, 294
method of infinite descent, 283	

Supplementary Exercises

1. What is the sum of an Eulerian and an Eisenstein number?
2. Is there only one zero or are there several? Could there be an infinite number of zeros?
3. Write a script that will list the primes of $\mathbb{Z}[\sqrt{-6}]$.
4. Explain why the Diophantine equation $x^2 + 61y^2 = 1$ can have only a finite number of solutions, whereas $x^2 - 61y^2 = 1$ has infinitely many solutions.

Chapter 13



THE ARITHMETIC OF IDEALS

SINCE $\mathbb{Z}[\sqrt{d}]$ does not, in general, have the property of unique factorization with respect to prime numbers, we added new primes with respect to which $\mathbb{Z}[\sqrt{d}]$ does have unique factorization.

13.1 Preliminaries

It would be natural at this point to go on and generalize the number systems of the previous chapter to the general form

$$\mathbb{Z}[\sqrt{d}] = \{x + y\sqrt{d} \mid x, y \in \mathbb{Z}\}$$

where d is a negative integer which is square-free. However, for pedagogical reasons, it is better to restrict attention to the systems in which $d \equiv 2, 3 \pmod{4}$ and d is square-free. The first seven such values of d are $-1, -2, -5, -6, -10, -13$, and -14 . For these values of d let

$$\mathbb{Z}[\sqrt{d}] = \{x + y\sqrt{d} \mid x, y \in \mathbb{Z}\}.$$

For example, when $d = -1$, $\mathbb{Z}[\sqrt{-1}]$ consists of all the Gaussian integers. The elements of $\mathbb{Z}[\sqrt{-2}]$ are, of course, the Eulerian integers, e.g., $1 - 2\sqrt{-2}$ and $3 + 5\sqrt{-2}$. It is clear that for these d 's that $\mathbb{Z}[\sqrt{d}] \subset \mathbb{C}$, that $\mathbb{Z}[\sqrt{d}] \not\subset \mathbb{R}$, and that $\mathbb{Z}[\sqrt{d}]$ is a subset of \mathbb{C} that is closed with respect to the operations of addition, subtraction, and multiplication. The reason for excluding the case $d \equiv 0 \pmod{4}$ is that these numbers are divisible by 4 and hence are not square-free. The reason for the exclusion of numbers that are not square-free is that when d is not square-free, say $d = ap^2$, then

$$x + y\sqrt{d} = x + py\sqrt{a} \in \mathbb{Z}[a]$$

so that $\mathbb{Z}[d] \subset \mathbb{Z}[a]$. Finally, the restriction $d \not\equiv 1 \pmod{4}$ is there to simplify the arguments of the main theorems. Recall that if $\alpha = x + y\sqrt{d}$ is any element of $\mathbb{Z}[\sqrt{d}]$, then the *conjugate* of α is $\bar{\alpha} = x - y\sqrt{d}$.

Proposition 13.1 Let $\alpha, \beta \in \mathbb{Z}[\sqrt{d}]$. Then

- (a) $\overline{\alpha + \beta} = \bar{\alpha} + \bar{\beta}$
- (b) $\overline{\alpha\beta} = \bar{\alpha}\bar{\beta}$
- (c) $\bar{\bar{\alpha}} = \alpha$
- (d) $\bar{\alpha} = \alpha$ if and only if $\alpha \in \mathbb{Z}$.

If $\alpha \in \mathbb{Z}[\sqrt{d}]$ then the *norm* of α is defined as $N(\alpha) = \alpha\bar{\alpha}$ and the *trace* of α is defined as $\text{Tr}(\alpha) = \alpha + \bar{\alpha}$. For example, if $\alpha = -2 - 5\sqrt{-3}$, then

$$N(\alpha) = (-2 - 5\sqrt{-3})(-2 + 5\sqrt{-3}) = (-2)^2 - 5^2(-3)^2 = 4 + 75 = 79$$

and

$$\text{Tr}(\alpha) = (-2 - 5\sqrt{-3}) + (-2 + 5\sqrt{-3}) = 2(-2) = -4.$$

Because d is negative, it is clear that if $\alpha = x + y\sqrt{d}$, then

$$N(\alpha) = x^2 - dy^2 > 0$$

and $N(\alpha)$ is a rational integer.

Proposition 13.2 If $\alpha \in \mathbb{Z}[\sqrt{d}]$, then $\text{Tr}(\alpha + \beta) = \text{Tr}(\alpha) + \text{Tr}(\beta)$ and $N(\alpha\beta) = N(\alpha)N(\beta)$.

Proof. See Exercise 13.1.1. ■

The set $\mathbb{Q}[\sqrt{d}] = \{x + y\sqrt{d} \mid x, y \in \mathbb{Q}\}$ is called a *quadratic field*. Fortunately, every quadratic field is indeed a field (see Exercise 13.1.3).

Proposition 13.3 Suppose $\kappa = \kappa_1 + \kappa_2\sqrt{d}$ where $\kappa_1, \kappa_2 \in \mathbb{Q}$ and $\text{Tr}(\kappa), N(\kappa) \in \mathbb{Z}$. Then $\kappa_1, \kappa_2 \in \mathbb{Z}$.

Proof. Let κ satisfy the hypotheses of this proposition. Since $\text{Tr}(\kappa) = 2\kappa_1$, κ_1 is either a rational integer or half an odd rational integer.

In the first case, the fact that $N(\kappa) = \kappa_1^2 - d\kappa_2^2 \in \mathbb{Z}$ implies that $d\kappa_2^2$ is also an integer. If m/n is the reduced form of κ_2 , then for some $z \in \mathbb{Z}$

$$\frac{dm^2}{n^2} = z \quad \text{or} \quad dm^2 = zn^2,$$

which, unless $n = \pm 1$, contradicts the fact that d is square-free. Hence κ_2 is the rational integer m . Thus, κ_1 and κ_2 are both in \mathbb{Z} .

In the second case, where κ_1 is half of an odd rational integer, set $\kappa_1 = a/2$ where a is an odd rational integer. Then, the norm of κ is

$$\kappa_1^2 - d\kappa_2^2 = \frac{a^2}{4} - d\kappa_2^2 \in \mathbb{Z}$$

and hence

$$a^2 - d(2\kappa_2)^2 \in 4\mathbb{Z}.$$

This, however, contradicts the fact that $a^2 \equiv 1 \pmod{4}$. ■

It is easy to see that $\mathbb{Z}[\sqrt{d}]$ is closed with respect to addition and subtraction. It calls for a little work to verify closure with respect to multiplication (see Exercise 13.1.2). As expected, the integers are not closed with respect to division. However, we do have the following useful fact regarding divisibility by rational integers.

Proposition 13.4 If $m \in \mathbb{Z}$ and $\alpha = a + b\sqrt{d}$, then $m \mid \alpha$ in $\mathbb{Z}[\sqrt{d}]$ if and only if $m \mid a$ and $m \mid b$ in \mathbb{Z} .

Proof. If $m \mid a$ and $m \mid b$ in \mathbb{Z} , then clearly $m \mid (a + b\sqrt{d})$. Conversely, suppose $m \mid (a + b\sqrt{d})$, meaning that there exists an integer $\alpha' = a' + b'\sqrt{d}$ such that

$$a + b\sqrt{d} = m(a' + b'\sqrt{d}) = ma' + mb'\sqrt{d}.$$

Since \sqrt{d} is a complex irrational number, it follows that $a = ma'$ and $b = mb'$. ■

Exercises 13.1

1. Prove Proposition 13.2.
2. Prove that $\mathbb{Z}[\sqrt{d}]$ is closed with respect to the operations of addition, subtraction, and multiplication.
3. Prove that the quadratic field $\mathbb{Q}[\sqrt{d}]$ is a field.

13.2 Integers of a Quadratic Field

In this section, a theory of factorization of integers is developed that is analogous to that which we know to hold for the rational integers. This “natural” approach is then

demonstrated to lead to undesirable consequences. In the previous chapter units were defined for \mathbb{Z} and $\mathbb{Z}[i]$; they are now extended to quadratic domains in general.

A unit of $\mathbb{Z}[\sqrt{-d}]$ (resp. \mathbb{Z}) is an element whose inverse is also in $\mathbb{Z}[\sqrt{-d}]$ (resp. \mathbb{Z}).

Proposition 13.5 An integer is a unit if and only if its norm is 1.

Proof. It is clear that the units of \mathbb{Z} are ± 1 and hence the proposition holds for the units of the rational integers.

Turning to $\mathbb{Z}[\sqrt{d}]$, let u and v both be units of $\mathbb{Z}[\sqrt{d}]$ such that $uv = 1$. The multiplicity property of the norm then yields $N(u)N(v) = 1$. Since the norm is a positive rational integer, it follows that $N(u) = 1$. If $d = -1$, then u must be one of the four numbers $\pm 1, \pm i$. If $d < -1$, then $N(x + y\sqrt{d}) = x^2 - dy^2$ which can only be 1 if $x = \pm 1$ and $y = 0$.

Conversely, ± 1 are clearly units of $\mathbb{Z}[\sqrt{d}]$, $d = -1, -2, -3, \dots$ and $\pm i$ are additional units of $\mathbb{Z}[\sqrt{-1}]$. ■

Corollary 13.6 The units of $\mathbb{Z}[\sqrt{-1}]$ are ± 1 and $\pm i$ whereas those of all other $\mathbb{Z}[\sqrt{d}]$ with negative d are ± 1 .

Let α and β be two integers such that for some unit u , $\alpha = \beta u$. Then α and β are said to be *associates* of each other. It follows from Corollary 13.6 that $a + bi$, $-b + ai$, $-a - bi$, and $b - ai$ are associates in \mathcal{I}_{-1} and $\alpha, -\alpha$ are each other's associates in $\mathbb{Z}[\sqrt{d}]$ for any $d = -1, -2, -3, \dots$

An element α of $\mathbb{Z}[\sqrt{d}]$ is *irreducible* if for any factorization $\alpha = \beta\gamma$ at least one of β and γ is necessarily a unit. The irreducible integers of \mathbb{Z} are the associates of the classical primes, and those of $\mathbb{Z}[i]$ are the Gaussian primes.

Theorem 13.7 If $\alpha \in \mathbb{Z}[\sqrt{d}]$ has a norm which is prime in \mathbb{Z} , then α is irreducible in $\mathbb{Z}[\sqrt{d}]$.

Proof. Let β and γ be integers in $\mathbb{Z}[d]$ such that $\alpha = \beta\gamma$. By Lemma 12.53

$$N(\alpha) = N(\beta)N(\gamma).$$

Since the norm of α is a prime in \mathbb{Z} , it follows that either $N(\beta)$ or $N(\gamma)$ equals 1 and so one of them is a unit. Hence α is irreducible. ■

For example, the integer $3 + 2\sqrt{-5}$ is irreducible in $\mathbb{Z}[\sqrt{-5}]$ because its norm is

$$3^2 - (-2)^2(-5) = 29.$$

The converse to Theorem 13.7 does not hold, as is demonstrated by the following examples:

In the Gaussian integers $\mathbb{Z}[\sqrt{-1}]$, $N(3) = 9$. Hence, if $3 = \alpha\beta$ is any factorization of 3 into Gaussian integers, both α and β must have norm 3. Since there are no such integers, $\mathbb{Z}[\sqrt{-1}]$, 3 is an irreducible integer whose norm is not prime in \mathbb{Z} .

In the ring¹ $\mathbb{Z}[\sqrt{-6}]$ the integer 2 is irreducible because the only integers whose norm properly divides $N(2) = 4$ are units. The same goes for 5.

In the ring $\mathbb{Z}[\sqrt{-6}]$, the integer $\alpha = 2 + \sqrt{-6}$ is irreducible. For, if $\alpha = \beta\gamma$, then $N(\beta) \mid N(\alpha)$. However, since there are no nonunits of $\mathbb{Z}[\sqrt{-6}]$ whose norms equal either 2 or 5, it follows that either α or β is a unit. Hence α is irreducible in $\mathbb{Z}[\sqrt{-6}]$. The same, of course, holds for $2 - \sqrt{-6}$.

Theorem 13.8 Every nonzero nonunit in $\mathbb{Z}[\sqrt{d}]$ is a product of irreducibles in $\mathbb{Z}[\sqrt{d}]$.

Proof. By induction on $N(\alpha)$. (See Theorem 4.9.) ■

For example, the integer 10 has two factorizations in $\mathbb{Z}[\sqrt{-6}]$:

$$2 \cdot 5 = 10 = (2 + \sqrt{-6})(2 - \sqrt{-6}).$$

Above, it was demonstrated that the integers 2, 5, and $2 \pm \sqrt{-6}$ are all irreducible in $\mathbb{Z}[\sqrt{-6}]$. Hence 10 has distinct factorizations into irreducible integers.

In 1847 the French mathematician Lamé famously announced that he had proven Fermat's "Last Theorem." His proof turned out to make the hidden assumption that if a perfect square is split into the product of two relatively prime integers, then each of the factors must itself be a square. This is easily enough proved to hold for the rational integers, but, as we now show, is false in general.

Consider the equation

$$2 \cdot (-3) = \sqrt{-6}^2.$$

Note that the right-hand side of this equation is a perfect square (i.e., the square of an integer). The two rational integers 2 and -3 are relatively prime because

$$1 = (-1)(2) + (-1)(-3)$$

¹At this point a *ring* is an algebraic structure with two binary operations resembling addition and multiplication. A more formal definition appears in the next chapter. Some of the better-known rings are \mathbb{Z} , \mathbb{Z}_n , $\mathbb{Z}[\sqrt{-6}]$, and $\mathbb{Z}[\sqrt{d}]$, as well as the rational, the real, and the complex numbers and the Galois fields of Chapter 7.

and hence their common factors are necessarily units. On the other hand, neither 2 nor -3 are squares nor are their associates. The reason for that is that the norm of any such “square root” must be either 2 or 3 and $\mathbb{Z}[\sqrt{-6}]$ contains no such integers.

Exercises 13.2

1. Find all the irreducible integers of $\mathbb{Z}[\sqrt{-5}]$ of norm less than 25.
2. Find all the irreducible integers of $\mathbb{Z}[\sqrt{-6}]$ of norm less than 25.
3. Factor all the integers of $\mathbb{Z}[\sqrt{-5}]$ of norm less than 30 into irreducible integers.
4. Factor all the integers of $\mathbb{Z}[\sqrt{-6}]$ of norm less than 30 into irreducible integers.

13.3 Ideals

By now the reader has seen enough examples to allow for the possibility that unique factorization might be the exception rather than the rule. In order to remedy this situation we introduce new entities called, for historical reasons, *ideals*. The four arithmetic operations are extended to these ideals, and eventually we demonstrate that these do have unique factorization into primes.

Let $A = \{\alpha_1, \alpha_2, \dots, \alpha_m\}$ be a finite set of elements of $\mathbb{Z}[\sqrt{d}]$. Then the (possibly infinite) set

$$\langle A \rangle = \langle \alpha_1, \alpha_2, \dots, \alpha_m \rangle = \left\{ r_1 \alpha_1 + r_2 \alpha_2 + \dots + r_m \alpha_m \mid r_1, r_2, \dots, r_m \in \mathbb{Z}[\sqrt{d}] \right\}$$

is called the *ideal of $\mathbb{Z}[\sqrt{d}]$ generated by A* .

For example, if $A = \{0\}$, then $\langle A \rangle = \{0\}$ for all d . If $A = \{1\}$, then $\langle A \rangle = \mathbb{Z}[\sqrt{d}]$ for all d . If $A = \{2\}$, then

$$\langle A \rangle = \left\{ a + b\sqrt{d} \mid a \text{ and } b \text{ are both even} \right\}.$$

If $A = \{2, 4\}$, then $\langle A \rangle = \langle \{2\} \rangle$. If $A = \{\sqrt{-5}, 3\}$, then $\langle A \rangle$ contains the following integers of $\mathbb{Z}[\sqrt{-5}]$:

$$0, 3, \sqrt{-5}, 3 - \sqrt{-5}, 4 + 3\sqrt{-5}, -15 + 4\sqrt{-5}$$

as well as

$$(1 + \sqrt{-5})3 + (2 - \sqrt{-5})\sqrt{-5} = 8 + 5\sqrt{-5}$$

and

$$(8 + 5\sqrt{-5})(1 - \sqrt{-5}) = 33 - 3\sqrt{-5}.$$

If $A = \{3, 2\}$, then $A = \langle 1 \rangle$. The reason for this is that $1 = 1 \cdot 3 + (-1) \cdot 2 \in \langle A \rangle$.

Proposition 13.9 Suppose α and β are elements of the ideal \mathfrak{a} of $\mathbb{Z}[\sqrt{d}]$. Then $\alpha + \beta \in \mathfrak{a}$ and $r\alpha \in \mathfrak{a}$.

Proof. By definition there exists a set $A = \{\alpha_1, \alpha_2, \dots, \alpha_m\}$ such that $\mathfrak{a} = \langle A \rangle$ and for each $i = 1, 2, \dots, m$ there exist integers r_i, s_i such that

$$\alpha = \sum_{i=1}^m r_i \alpha_i \quad \text{and} \quad \beta = \sum_{i=1}^m s_i \alpha_i.$$

Then

$$\alpha + \beta = \sum_{i=1}^m r_i \alpha_i + \sum_{i=1}^m s_i \alpha_i = \sum_{i=1}^m (r_i + s_i) \alpha_i \in \mathfrak{a}.$$

Moreover

$$r\alpha = r \sum_{i=1}^m r_i \alpha_i = \sum_{i=1}^m r r_i \alpha_i \in \mathfrak{a}. \quad \blacksquare$$

Proposition 13.10 If u is a unit of $\mathbb{Z}[\sqrt{d}]$, then $\langle u \rangle = \mathbb{Z}[\sqrt{d}]$.

Proof. Since u is a unit there exists an integer v such that $uv = 1$. By Proposition 13.10, since $u \in \mathfrak{a}$, so is $uv = 1$ in \mathfrak{a} , and hence $1 \in \mathfrak{a}$. Then, however, by the same proposition, for any α in \mathfrak{a} , $\alpha = \alpha \cdot 1 \in \mathfrak{a}$. \blacksquare

If A is a singleton, say $A = \{\alpha\}$, then $\langle A \rangle = \langle \alpha \rangle$ is said to be a *principal ideal*. Clearly $\langle 2 \rangle$ and $\langle 102 \rangle$ are principal ideals of \mathbb{Z} and $\langle 2 - 237\sqrt{-6} \rangle$ is an ideal of $\mathbb{Z}[\sqrt{-6}]$. We give an example of an ideal which is not principal: Consider the ideal $I = \langle 2, \sqrt{-6} \rangle$ of $\mathbb{Z}[\sqrt{-6}]$ and suppose that there is an integer α of $\mathbb{Z}[\sqrt{-6}]$ such that $I = \langle \alpha \rangle$. We first note that the norm of every element of I is even. This is justified as follows. By definition, every integer in I can be expressed in the form

$$(a + b\sqrt{-6}) \cdot 2 + (a' + b'\sqrt{-6}) \cdot \sqrt{-6} = (2a - 6b') + (2b + a')\sqrt{-6}.$$

This integer has norm

$$(2a - 6b')^2 + 6(2b + a')^2$$

which is clearly even.

Since $2 \in I = \langle \alpha \rangle$, it follows that there is an integer $\beta \in \mathbb{Z}[\sqrt{-6}]$ such that $2 = \alpha\beta$. Since the norm is multiplicative,

$$4 = N(2) = N(\alpha)N(\beta),$$

and we may conclude that $N(\alpha) = 1$ or 2 . However, no integer in $\mathbb{Z}[\sqrt{-6}]$ has norm 2 (see Exercise 13.2.2) so that necessarily $N(\alpha) = 1$. It follows that α is a unit and hence

$$I = \langle \alpha \rangle = \langle 1 \rangle.$$

This means that $1 \in I$, contradicting the fact that all the elements of I have even norm. Thus, the ideal $I = \langle 2, \sqrt{-6} \rangle$ is not principal.

Proposition 13.11 Let $\mathfrak{a} = \langle \alpha_1, \alpha_2, \dots, \alpha_m \rangle$ and $\mathfrak{b} = \langle \beta_1, \beta_2, \dots, \beta_n \rangle$ be two ideals in $\mathbb{Z}[\sqrt{d}]$. Then the following are equivalent:

- (a) $\mathfrak{a} \subset \mathfrak{b}$;
- (b) each α_i is in \mathfrak{b} ;
- (c) each α_i is a $\mathbb{Z}[\sqrt{d}]$ -linear combination of the β_j 's.

Proof. (a) implies (b) is clear and that (b) implies (c) follows from the definition of ideals. For $c \Rightarrow a$, Let α be any integer of \mathfrak{a} . This means that α is a $\mathbb{Z}[\sqrt{d}]$ -linear combination of the α_i 's. By hypothesis, each α_i is a $\mathbb{Z}[\sqrt{d}]$ -linear combination of the β_j 's. It follows that α is a $\mathbb{Z}[\sqrt{d}]$ -linear combination of the β_j 's. Hence, $\alpha \in \mathfrak{b}$ as well. ■

Corollary 13.12 Suppose $A = \alpha_1, \alpha_2, \dots, \alpha_m$ and $B = \beta_1, \beta_2, \dots, \beta_n$. Then $\langle A \rangle = \langle B \rangle$ if and only if $A \subset \langle B \rangle$ and $B \subset \langle A \rangle$.

The next corollary states that the addition of a multiple of one generator to another does not affect the span.

Corollary 13.13 If $\alpha_1, \alpha_2, \dots, \alpha_m, \gamma \in \mathbb{Z}[\sqrt{d}]$, then

$$\langle \alpha_1, \alpha_2, \dots, \alpha_m \rangle = \langle \alpha_1 + \gamma\alpha_2, \alpha_2, \dots, \alpha_m \rangle.$$

For example, to simplify

$$\langle 14 + 9\sqrt{-6}, 16 - 3\sqrt{-6}, 2 + 6\sqrt{-6}, 10 \rangle,$$

subtract twice the second generator from the third one to obtain

$$\langle 14 + 9\sqrt{-6}, 16 - 3\sqrt{-6}, -34, 10 \rangle.$$

Next add three times the second generator to the first one, resulting in

$$\langle 62, 16 - 3\sqrt{-6}, -34, 10 \rangle$$

which has the same span as $\langle 2, 3\sqrt{-6} \rangle$.

Proposition 13.14 Every ideal of \mathbb{Z} is principal.

Proof. Let d be the least positive element of the ideal I of \mathbb{Z} . If a is any integer of I there exist an integer q and a nonnegative integer $r < d$ such that $a = qd + r$. Since $a, d \in I$, it follows that $r = a - qd \in I$. The minimality of d implies that $r = 0$ from which it follows that $a = qd \in \langle d \rangle$, and hence $I = \langle d \rangle$. ■

Proposition 13.14 above implies that in order to find a nonprincipal ideal we must look elsewhere than \mathbb{Z} and we now describe a nonprincipal ideal in $\mathbb{Z}[\sqrt{-3}]$:

$$I = \langle 2, 1 + \sqrt{-3} \rangle.$$

If this ideal were principal, say $I = \langle z \rangle$, the generator z would have to be a common divisor of both 2 and $1 + \sqrt{-3}$, both of which are prime (see Table 12.6). It follows that z would have to be a unit, i.e., ± 1 . It is easy to see that $1 \in I$ if and only if $-1 \in I$ and hence it suffices to show that $1 \notin I$. Suppose, by way of contradiction, that $1 \in I$. Hence, there exist rational integers x, y, u, v such that

$$1 = 2(x + y\sqrt{-3}) + (1 + \sqrt{-3})(u + v\sqrt{-3}).$$

Equating rational and complex terms we get

$$1 = 2x + u - 3v$$

$$0 = 2y + v + u.$$

The subtraction of these two equations yields $1 = 2x - 2y - 4v$, which is impossible since x, y, v are all rational integers. Hence $\langle 2, 1 + \sqrt{-3} \rangle$ is a nonprincipal ideal of $\mathbb{Z}[\sqrt{-3}]$.

We continue this exposition with the following observations about the ring of rational integers.

Proposition 13.15 The correspondence $a \leftrightarrow \langle a \rangle$ between the nonnegative integers of \mathbb{Z} and its ideals is a bijection. Moreover $a \mid b$ if and only if $\langle a \rangle \supseteq \langle b \rangle$.

Proof. Clearly, if $a = b$, then $\langle a \rangle = \langle b \rangle$. If $\langle a \rangle = \langle b \rangle$, then $a = \pm b$; since both a and b are positive, $a = b$. This takes care of all the ideals of \mathbb{Z} for the first part.

If $a \mid b$, then there exists an integer c such that $b = ac$. Consequently $b \in \langle a \rangle$ and hence $\langle a \rangle \supseteq \langle b \rangle$. Conversely, if $\langle a \rangle \supseteq \langle b \rangle$ then $b \in \langle a \rangle$ and hence there exists an integer c such that $ac = b$, i.e., a divides b . ■

If $A = \{a_1, a_2, \dots, a_s\}$ and $B = \{b_1, b_2, \dots, b_t\}$, then the *product* $A \cdot B$ or AB is defined as

$$A \cdot B = AB = \langle a_1 b_1, a_1 b_2, \dots, a_1 b_t, a_2 b_1, a_2 b_2, \dots, a_2 b_t, \dots, a_s b_1, a_s b_2, \dots, a_s b_t \rangle.$$

For example,

$$\begin{aligned} \langle 2, 1 + \sqrt{-5} \rangle^2 &= \langle 2, 1 + \sqrt{-5} \rangle \langle 2, 1 + \sqrt{-5} \rangle \\ &= \langle 4, 2 + 2\sqrt{-5}, -4 + 2\sqrt{-5} \rangle = \langle 4, 2 + 2\sqrt{-5}, -4 + 2\sqrt{-5}, 2 \rangle = \langle 2 \rangle \end{aligned}$$

because each of the generators of the penultimate expression is divisible by 2 and, moreover, 2 is one of its generators. Also,

$$\begin{aligned} \langle 3, 1 + \sqrt{-5} \rangle \langle 3, 1 - \sqrt{-5} \rangle &= \langle 9, 3 - 3\sqrt{-5}, 3 + 3\sqrt{-5}, 6 \rangle \\ &= \langle 9, 3 - 3\sqrt{-5}, 3 + 3\sqrt{-5}, 6, 3 \rangle = \langle 3 \rangle \end{aligned}$$

because each of the generators of the penultimate expression is divisible by 3 and, moreover, 3 is one of its generators.

Proposition 13.16 If $\langle a \rangle$ and $\langle b \rangle$ are two elements of the ring R , then $\langle a \rangle \langle b \rangle = \langle ab \rangle$.

Proof. This follows directly from the definition of the multiplication of ideals. ■

Let $S = \{a_1, a_2, \dots, a_s\}$ and $T = \{b_1, b_2, \dots, b_t\}$ be two arbitrary finite subsets of a ring. The two ideals $A = \langle S \rangle$ and $B = \langle T \rangle$ are said to be *equal* if they are equal as sets.

Equivalently, they are equal if $S \subseteq B$ and $T \subseteq A$. It is clear that permuting the order of the generators does not alter the ideal they determine. Moreover, if $G = \{g_1, g_2, \dots, g_b\}$, $G' = \{g_1, g_2, \dots, g_{b-1}\}$, and $g_b \in \langle G' \rangle$, then $\langle G \rangle = \langle G' \rangle$, in which case we say that $\langle G' \rangle$ is obtained from $\langle G \rangle$ by the *removal* of g_b and $\langle G \rangle$ is obtained from $\langle G' \rangle$ by the *addition* of g_b .

For example, $\langle 2, 1 - \sqrt{-7} \rangle = \langle 2, 1 + \sqrt{-7} \rangle$ is demonstrated by the following chain of equations:

$$\langle 2, 1 - \sqrt{-7} \rangle = \langle 2, 1 - \sqrt{-7}, 1 + \sqrt{-7} \rangle = \langle 2, 1 + \sqrt{-7} \rangle.$$

To see that $\langle 2, 1 + \sqrt{-5} \rangle \langle 3, 1 + \sqrt{-5} \rangle = \langle 1 + \sqrt{-5} \rangle$, compute

$$\begin{aligned} \langle 2, 1 + \sqrt{-5} \rangle \langle 3, 1 + \sqrt{-5} \rangle &= \langle 6, 2(1 + \sqrt{-5}), 3(1 + \sqrt{-5}), 1 + 2\sqrt{-5} - 5 \rangle \\ &= \langle 6, 2 + 2\sqrt{-5}, 3 + 3\sqrt{-5} \rangle \\ &= \langle 1 + \sqrt{-5} \rangle \end{aligned}$$

since $6 = (1 + \sqrt{-5})(1 - \sqrt{-5})$.

To see that $\langle 2, 1 + \sqrt{-5} \rangle \langle 3, 1 - \sqrt{-5} \rangle = \langle 1 - \sqrt{-5} \rangle$, compute

$$\begin{aligned} \langle 2, 1 + \sqrt{-5} \rangle \langle 3, 1 - \sqrt{-5} \rangle &= \langle 6, 2 - 2\sqrt{-5}, 3 + 3\sqrt{-5}, 6 \rangle \\ &= \langle 6, 2 - 2\sqrt{-5}, 3 + 3\sqrt{-5} \rangle \\ &= \langle 6 - (2 - 2\sqrt{-5} - (3 + 3\sqrt{-5})), 2 - 2\sqrt{-5}, 3 + 3\sqrt{-5} \rangle \\ &= \langle 1 - \sqrt{-5}, 2 - 2\sqrt{-5}, 3 + 3\sqrt{-5} \rangle \\ &= \langle 1 - \sqrt{-5} \rangle. \end{aligned}$$

Proposition 13.17 If \mathfrak{a} , \mathfrak{b} , and \mathfrak{c} are ideals of $\mathbb{Z}[\sqrt{d}]$, then $\mathfrak{a}\mathfrak{b} = \mathfrak{b}\mathfrak{a}$ and $(\mathfrak{a}\mathfrak{b})\mathfrak{c} = \mathfrak{a}(\mathfrak{b}\mathfrak{c})$.

Proof. Let $\mathfrak{a} = \langle \alpha_1, \alpha_2, \dots, \alpha_m \rangle$, $\mathfrak{b} = \langle \beta_1, \beta_2, \dots, \beta_n \rangle$, and $\mathfrak{c} = \langle \gamma_1, \gamma_2, \dots, \gamma_p \rangle$. Then

$$\mathfrak{a}\mathfrak{b} = \langle \dots, \alpha_i \beta_j, \dots \rangle = \langle \dots, \beta_j \alpha_i, \dots \rangle = \mathfrak{b}\mathfrak{a}$$

and

$$(\mathfrak{a}\mathfrak{b})\mathfrak{c} = \langle \dots, (\alpha_i \beta_j) \gamma_k, \dots \rangle = \langle \dots, \alpha_i (\beta_j \gamma_k), \dots \rangle = \mathfrak{a}(\mathfrak{b}\mathfrak{c}). \quad \blacksquare$$

For example,

$$\begin{aligned}\langle 3, 1 + \sqrt{-5} \rangle \langle 3, 1 - \sqrt{-5} \rangle &= \langle 9, 3 - 3\sqrt{-5}, 3 + 3\sqrt{-5}, 6 \rangle \\ &= \langle 9, 3 - 3\sqrt{-5}, 3 + 3\sqrt{-5}, 6, 3 \rangle = \langle 3 \rangle.\end{aligned}$$

In $\mathbb{Z}[\sqrt{-6}]$

$$\langle 2, 1 + \sqrt{-6} \rangle = \langle 1 \rangle. \quad (13.18)$$

To see this, note that

$$7 = (1 + \sqrt{-6})(1 - \sqrt{-6}) \in \langle 2, 1 + \sqrt{-6} \rangle.$$

Hence 7 and 2 are also in that same ideal. However, since 7 and 2 are relatively prime in \mathbb{Z} , there exist rational integers A and B such that $A \cdot 7 + B \cdot 2 = 1$, and hence Equation 13.18 holds.

In $\mathbb{Z}[\sqrt{-6}]$

$$\langle 2 + \sqrt{-6}, 7 + 2\sqrt{-6} \rangle = \langle 3, 1 - \sqrt{-6} \rangle.$$

This follows from the observation that each generator of one ideal is a $\mathbb{Z}[\sqrt{-6}]$ -linear combination of the other ideal:

$$3 = (7 + 2\sqrt{-6}) - 2(2 + \sqrt{-6}) \quad \text{and} \quad 1 - \sqrt{-6} = (7 + 2\sqrt{-6}) - 3(2 + \sqrt{-6}),$$

and, in the other direction,

$$2 + \sqrt{-6} = 3 - (1 - \sqrt{-6}) \quad \text{and} \quad 7 + 2\sqrt{-6} = 3(1 + 2\sqrt{-6}) + 4(1 - \sqrt{-6}).$$

Could the ideal in question be $\langle 1 \rangle$? (See Exercise 13.3.8.)

In $\mathbb{Z}[\sqrt{-6}]$,

$$\langle 4 + \sqrt{-6}, 2 - \sqrt{-6}, 7 + \sqrt{-6} \rangle = \langle 3, 1 + \sqrt{-6} \rangle.$$

This is proved by exhibiting each generator of one ideal as a $\mathbb{Z}[\sqrt{-6}]$ -linear combination of the generators of the other ideal:

$$3 = -2(2 - \sqrt{-6}) + (7 - 2\sqrt{-6}) \quad \text{and} \quad 1 + \sqrt{-6} = 2(4 + \sqrt{-6}) - (7 + \sqrt{-6}),$$

and, in the opposite direction,

$$4 + \sqrt{-6} = 1 \cdot 3 + 1 \cdot (1 + \sqrt{-6}), \quad 2 - \sqrt{-6} = 1 \cdot 3 - 1 \cdot (1 + \sqrt{-6}),$$

and

$$7 + \sqrt{-6} = 2 \cdot 3 + 1 \cdot (1 + \sqrt{-6}).$$

Proposition 13.19 If an ideal in $\mathbb{Z}[\sqrt{d}]$ contains two elements of \mathbb{Z} which are relatively prime, then the ideal is the unit ideal. Consequently, an ideal is the unit ideal if it contains two elements whose norms are relatively prime in \mathbb{Z} .

Proof. Let α and β be elements of the ideal I which are also in \mathbb{Z} and are relatively prime there. Then there exist rational integers A and B such that $1 = A\alpha + B\beta$. On the other hand every integer of the form $A\alpha + B\beta$ is contained in I . Hence $1 \in I$. By the defining properties of ideals $A\alpha + B\beta \in I$, so it follows that 1 is also in I , ergo I is the unit ideal.

Suppose next that $\alpha, \beta \in I$ are such that $N(\alpha)$ and $N(\beta)$ are relatively prime. Since $N(\alpha) = \alpha\bar{\alpha} \in I$ and similarly $N(\beta) \in I$, it then follows from the first half of the proof first that 1 is also in I and second that I is the unit ideal. ■

Proposition 13.20 If all the generators of the ideal I are rational integers, then I is the principal ideal $\langle h \rangle$ where h is the greatest common divisor of the generators.

Proof. Suppose $I = \langle a_1, a_2, \dots, a_m \rangle$ and let h be the greatest common divisor of all the a_i 's (which are all in \mathbb{Z}). It follows that $h \mid a_i$ for each $i = 1, 2, \dots, m$, so that

$$\{a_1, a_2, \dots, a_m\} \subset \langle h \rangle$$

and hence $I \subset \langle h \rangle$. To obtain the reverse inclusion, note that by Theorem 4.8 there exist rational integers A_1, A_2, \dots, A_m such that

$$h = A_1 a_1 + A_2 a_2 + \dots + A_m a_m.$$

Consequently, for any integer α

$$\alpha h = (\alpha A_1) a_1 + (\alpha A_2) a_2 + \dots + (\alpha A_m) a_m \in I$$

and hence $\langle h \rangle \subset I$. ■

Proposition 13.21 Let α and β be integers of the ideal I . Then $\langle \alpha \rangle = \langle \beta \rangle$ if and only if α and β are associates.

Proof. If one of α and β is 0, we may assume without loss of generality that $\beta = 0$. In that case the following are equivalent:

- $\langle \alpha \rangle = \langle \beta \rangle$;
- $\langle \alpha \rangle = \langle 0 \rangle$;
- $\langle \alpha \rangle = \{0\}$;
- $\alpha = 0$;
- $\alpha = \beta$.

Thus, we may assume that neither α nor β is zero. By Proposition 13.11, the following are equivalent

- $\langle \alpha \rangle = \langle \beta \rangle$;
- $\alpha \in \langle \beta \rangle$ and $\beta \in \langle \alpha \rangle$;
- $\alpha = \beta\gamma$ and $\beta = \alpha\gamma'$ for some $\gamma, \gamma' \in \mathbb{Z}[\sqrt{d}]$.

Consequently, if $\langle \alpha \rangle = \langle \beta \rangle$, then $\alpha = \beta\gamma = \alpha\gamma'\gamma$ from which it follows that $\gamma'\gamma = 1$ so that both γ and γ' are units.

Conversely, if α and β are associates, then $\alpha \mid \beta$ and $\beta \mid \alpha$ and hence $\langle \alpha \rangle = \langle \beta \rangle$. ■

Suppose \mathfrak{a} and \mathfrak{b} are two ideals of $\mathbb{Z}[\sqrt{d}]$. Their product \mathfrak{ab} is the set of all finite sums of the form

$$\left\{ \sum_{k=1}^m x_k y_k \mid x_k \in \mathfrak{a}, y_k \in \mathfrak{b}, i = 1, 2, \dots, m \right\}.$$

The rationale behind this definition is that at the very least the “product” of \mathfrak{a} and \mathfrak{b} should contain all products of the form xy where $x \in \mathfrak{a}$ and $y \in \mathfrak{b}$. However, this product should also be an ideal and so it must be closed under addition. Hence the summation.

Proposition 13.22 If $\mathfrak{a} = \langle \alpha_1, \alpha_2, \dots, \alpha_m \rangle$ and $\mathfrak{b} = \langle \beta_1, \beta_2, \dots, \beta_n \rangle$, then

$$\mathfrak{ab} = \langle \alpha_1\beta_1, \dots, \alpha_i\beta_j, \dots, \alpha_m\beta_n \rangle.$$

Proof. Let $\sum_{k=1}^m x_k y_k$ be an arbitrary element of \mathfrak{ab} . Note that each x_k is a $\mathbb{Z}[\sqrt{d}]$ -linear combination of the α_i 's and each y_k is a $\mathbb{Z}[\sqrt{d}]$ -linear combination of the β_j 's. When these multiplications are carried out, we obtain $\sum_{k=1}^m x_k y_k$ as a $\mathbb{Z}[\sqrt{d}]$ -linear combination of the $\alpha_i\beta_j$'s. Hence

$$\mathfrak{ab} \subset \langle \alpha_1\beta_1, \dots, \alpha_i\beta_j, \dots, \alpha_m\beta_n \rangle.$$

Conversely, every element of $\langle \alpha_1\beta_1, \dots, \alpha_i\beta_j, \dots, \alpha_m\beta_n \rangle$ can be written in the form

$$\sum_{i=1}^m \sum_{j=1}^n \gamma_{ij} \alpha_i \beta_j. \quad (13.23)$$

Since $\gamma_{ij}\alpha_i \in \mathfrak{a}$ and $\beta_j \in \mathfrak{b}$, it follows that the sum of Equation 13.23 is in \mathfrak{ab} . ■

Consider the product of the ideal $\mathfrak{a} = \langle 5 + \sqrt{-6}, 2 + \sqrt{-6} \rangle$ and the ideal $\mathfrak{b} = \langle 4 + \sqrt{-6}, 2 - \sqrt{-6} \rangle$. By definition, \mathfrak{ab} is generated by the products $(5 + \sqrt{-6})(3 + \sqrt{-6})$, $(5 + \sqrt{-6})(2 - \sqrt{-6})$, $(2 + \sqrt{-6})(4 + \sqrt{-6})$, and $(2 + \sqrt{-6})(2 - \sqrt{-6})$, so that

$$\mathfrak{ab} = \langle 14 + 9\sqrt{-6}, 16 - 3\sqrt{-6}, 2 + 6\sqrt{-6}, 10 \rangle,$$

which, according to an earlier example, simplifies to $\langle 2, 3\sqrt{-6} \rangle$.

Corollary 13.24 For any two ideals \mathfrak{a} and \mathfrak{b} , $\mathfrak{ab} = \langle 0 \rangle$ if and only if $\mathfrak{a} = \langle 0 \rangle$ or $\mathfrak{b} = \langle 0 \rangle$.

Proof. It follows from the definition that if either \mathfrak{a} or \mathfrak{b} is $\langle 0 \rangle$, then $\mathfrak{ab} = \langle 0 \rangle$. Conversely, suppose $\mathfrak{ab} = \langle 0 \rangle$ but both \mathfrak{a} and \mathfrak{b} are nonzero. Then there exist nonzero integers $\alpha \in \mathfrak{a}$ and $\beta \in \mathfrak{b}$. Hence $\alpha\beta$ is a nonzero element of \mathfrak{ab} , contradicting the fact that $\mathfrak{ab} = \langle 0 \rangle$. ■

Corollary 13.25 For the ideal $\mathfrak{a} = \langle \alpha_1, \alpha_2, \dots, \alpha_m \rangle$ and the principal ideal $\mathfrak{c} = \mathfrak{a}\langle \gamma \rangle$,

$$\mathfrak{a}\gamma = \langle \gamma\alpha_1, \gamma\alpha_2, \dots, \gamma\alpha_m \rangle.$$

In particular, the unit ideal $\langle 1 \rangle = \mathbb{Z}[\sqrt{d}]$ is an identity element for the multiplication of ideals, and $\langle \alpha, \beta \rangle = \langle \alpha \rangle \langle \beta \rangle$.

For an ideal \mathfrak{a} , its *conjugate ideal* is the ideal $\bar{\mathfrak{a}} = \{ \bar{\alpha} \mid \alpha \in \mathfrak{a} \}$.

Proposition 13.26 If $\mathfrak{a} = \langle \alpha_1, \alpha_2, \dots, \alpha_m \rangle$, then $\bar{\mathfrak{a}} = \langle \bar{\alpha}_1, \bar{\alpha}_2, \dots, \bar{\alpha}_m \rangle$. In particular, if $\mathfrak{a} = \langle \alpha \rangle$ is a principal ideal, then so is $\bar{\mathfrak{a}}$ the principal ideal $\langle \bar{\alpha} \rangle$.

Proof. By definition,

$$\mathfrak{a} = \left\{ \sum_{i=1}^m r_i \alpha_i \mid r_1, \dots, r_m \in \mathbb{Z}[\sqrt{d}] \right\}.$$

Making use of the additivity and multiplicativity of conjugation we obtain

$$\begin{aligned}\bar{\mathfrak{a}} &= \left\{ \sum_{i=1}^m \bar{r}_i \bar{\alpha}_i \mid r_1, \dots, r_m \in \mathbb{Z}[\sqrt{d}] \right\} \\ &= \left\{ \sum_{i=1}^m \bar{r}_i \bar{\alpha}_i \mid \bar{r}_1, \dots, \bar{r}_m \in \mathbb{Z}[\sqrt{d}] \right\} \\ &= \langle \bar{\alpha}_1, \bar{\alpha}_2, \dots, \bar{\alpha}_m \rangle.\end{aligned}$$

The remainder of this proposition is relegated to Exercise 13.3.6. ■

When an integer is equal to its conjugate, they must both be in \mathbb{Z} . This is not the case for self-conjugate ideals. Consider the ideal $\langle 2, \sqrt{-6} \rangle$. Its conjugate is

$$\overline{\langle 2, \sqrt{-6} \rangle} = \langle \bar{2}, \bar{\sqrt{-6}} \rangle = \langle 2, -\sqrt{-6} \rangle$$

and so we see that even though the ideal is self-conjugate, it need not consist of rational integers only.

The ideal $\langle 5, 2 + \sqrt{-6} \rangle$ is not principal and is not its own conjugate. We first show that

$$\langle 5, 2 + \sqrt{-6} \rangle \langle 5, 2 - \sqrt{-6} \rangle = \langle 5 \rangle. \quad (13.27)$$

Indeed,

$$\begin{aligned}\langle 5, 2 + \sqrt{-6} \rangle \langle 5, 2 - \sqrt{-6} \rangle &= \langle 25, 10 - 5\sqrt{-6}, 10 + 5\sqrt{-6}, 4 - (-6) \rangle \\ &= \langle 5 \rangle \langle 5, 2 - \sqrt{-6}, 2 + \sqrt{-6}, 2 \rangle \\ &= \langle 5 \rangle \langle 1 \rangle = \langle 5 \rangle.\end{aligned}$$

Now, suppose by way of contradiction that the given ideal is principal. Then there exists an integer α such that $\langle 5, 2 + \sqrt{-6} \rangle = \langle \alpha \rangle$. Hence $\langle \alpha \rangle \langle \bar{\alpha} \rangle = \langle 5 \rangle$. Thus,

$$\langle 5 \rangle = \langle \alpha \rangle \langle \bar{\alpha} \rangle = \langle \alpha \bar{\alpha} \rangle = \langle N(\alpha) \rangle$$

and it follows that $N(\alpha) = \pm 5$. Since $\mathbb{Z}[\sqrt{-6}]$ contains no elements of norm 5, we have our contradiction.

To show that the given ideal is not equal to its conjugate, we suppose, again by way of contradiction, that it does equal its conjugate ideal. Then Equation 13.27 becomes

$$\langle 5, 2 + \sqrt{-6} \rangle^2 = \langle 5 \rangle.$$

When we compute the above squared ideal $\langle 5, 2 + \sqrt{-6} \rangle^2$ directly, we obtain

$$\begin{aligned} \langle 5 \rangle &= \langle 5, 2 + \sqrt{-6} \rangle^2 = \langle 5, 2 + \sqrt{-6} \rangle \langle 5, 2 + \sqrt{-6} \rangle \\ &= \langle 25, 5(2 + \sqrt{-6}), (2 + \sqrt{-6})^2 \rangle \\ &= \langle 25, 5(2 + \sqrt{-6}), 4 + 4\sqrt{-6} - 6 \rangle \\ &= \langle 25, 5(2 + \sqrt{-6}), -2 + 4\sqrt{-6} \rangle. \end{aligned}$$

However, 5 is not an integer factor of $-2 + 4\sqrt{-6}$ and this is the required contradiction.

Let

$$p = \langle 5, 2 + \sqrt{-6} \rangle \quad \text{and} \quad q = \langle 2, 2 + \sqrt{-6} \rangle$$

so that

$$\bar{p} = \langle 5, 2 - \sqrt{-6} \rangle \quad \text{and} \quad \bar{q} = \langle 2, 2 - \sqrt{-6} \rangle.$$

Then it can be demonstrated (see Exercise 13.3.7) that

$$p\bar{p} = \langle 5 \rangle, \quad q\bar{q} = \langle 2 \rangle, \quad \text{and} \quad pq = \langle 2 + \sqrt{-6} \rangle. \quad (13.28)$$

Set $a \mid b$ if $b = ac$ for some ideal c . We say that a *divides* b or is a *divisor* or a *factor* of b , or that b is a *multiple* of a .

For example, since $\langle 2 \rangle \langle 3 \rangle = \langle 6 \rangle$, it follows that $\langle 2 \rangle \mid \langle 6 \rangle$ and $\langle 3 \rangle \mid \langle 6 \rangle$. Similarly, since

$$(1 + 2\sqrt{-6})(1 + \sqrt{-6}) = -11 + 3\sqrt{-6}$$

it follows that

$$(1 + 2\sqrt{-6}) \mid \langle -11 + 3\sqrt{-6} \rangle \quad \text{and} \quad \langle 1 + \sqrt{-6} \rangle \mid \langle -11 + 3\sqrt{-6} \rangle.$$

Proposition 13.29 If $\alpha, \beta \in \mathbb{Z}[\sqrt{d}]$, then $\langle \alpha \rangle \mid \langle \beta \rangle$ if and only if $\alpha \mid \beta$.

Proof. Suppose $\alpha \mid \beta$ in $\mathbb{Z}[\sqrt{d}]$. This means that there exists $\gamma \in \mathbb{Z}[\sqrt{d}]$ such that $\beta = \alpha\gamma$. Then, by the above proposition $\langle \beta \rangle = \langle \alpha\gamma \rangle = \langle \alpha \rangle \langle \gamma \rangle$. Hence $\langle \alpha \rangle \mid \langle \beta \rangle$. Conversely,

assume that $\langle \alpha \rangle \mid \langle \beta \rangle$. Then there exists an ideal \mathfrak{c} such that $\langle \beta \rangle = \langle \alpha \rangle \mathfrak{c}$. By definition, there exist integers $\gamma_1, \gamma_2, \dots, \gamma_p$ such that

$$\mathfrak{c} = \langle \gamma_1, \gamma_2, \dots, \gamma_p \rangle.$$

By Corollary 13.25

$$\langle \beta \rangle = \langle \alpha \rangle \langle \gamma_1, \gamma_2, \dots, \gamma_p \rangle = \langle \alpha\gamma_1, \alpha\gamma_2, \dots, \alpha\gamma_p \rangle.$$

Hence, like all the elements of $\langle \beta \rangle$, the integer β is a $\mathbb{Z}[\sqrt{d}]$ -linear combination of the $\alpha\gamma_k$'s, say

$$\beta = \sum_{k=1}^p \delta_k \alpha\gamma_k = \alpha \sum_{k=1}^p \delta_k \gamma_k.$$

Consequently $\alpha \mid \beta$. ■

Proposition 13.30 Let α be an integer and $\mathfrak{b} = \langle \beta_1, \beta_2, \dots, \beta_n \rangle$ an ideal in $\mathbb{Z}[\sqrt{d}]$. Then the following are equivalent:

- (a) $\langle \alpha \rangle \mid \mathfrak{b}$;
- (b) $\alpha \mid \beta_j$ for all $j = 1, 2, \dots, n$;
- (c) $\langle \alpha \rangle \supset \mathfrak{b}$.

Proof. (a) \Rightarrow (b): Suppose $\langle \alpha \rangle \mid \mathfrak{b}$. This means that there is an ideal $\mathfrak{c} = \langle \gamma_1, \gamma_2, \dots, \gamma_k \rangle$ such that

$$\mathfrak{b} = \langle \alpha \rangle \mathfrak{c} = \langle \alpha\gamma_1, \alpha\gamma_2, \dots, \alpha\gamma_p \rangle.$$

Hence every element of \mathfrak{b} is divisible by α , including the β_j 's.

(b) \Rightarrow (c): Suppose $\alpha \mid \beta_j$ for all $j = 1, 2, \dots, n$. Then clearly α must divide all the elements of $\langle \beta_1, \beta_2, \dots, \beta_n \rangle$. Consequently, since $\langle \alpha \rangle$ consists of all the integers divisible by α , we have $\langle \alpha \rangle \supset \mathfrak{b}$.

(c) \Rightarrow (a): Suppose $\langle \alpha \rangle \supset \mathfrak{b}$. Then every integer of \mathfrak{b} is a multiple of α . The set

$$\mathfrak{b}/\alpha = \left\{ m/\alpha \mid m \in \mathfrak{b} \right\}$$

is an ideal with generating set $\beta_1/\alpha, \beta_2/\alpha, \dots, \beta_n/\alpha$. By the definition of the product of ideals, $\langle \alpha \rangle (\mathfrak{b}/\alpha) = \mathfrak{b}$, and hence $\langle \alpha \rangle \mid \mathfrak{b}$. ■

Proposition 13.31 Let \mathfrak{a} and \mathfrak{b} be two ideals of $\mathbb{Z}[\sqrt{d}]$. Then

- (a) if $\mathfrak{a} \mid \mathfrak{b}$, then $\mathfrak{a} \supset \mathfrak{b}$;
- (b) if $\mathfrak{a} \mid \mathfrak{b}$ and $\mathfrak{b} \mid \mathfrak{a}$, then $\mathfrak{a} = \mathfrak{b}$.

Proof. Since (b) follows immediately from (a), we only prove part (a): If $\mathfrak{a} \mid \mathfrak{b}$, then there exists an ideal \mathfrak{c} such that $\mathfrak{b} = \mathfrak{a}\mathfrak{c}$. Suppose that $\mathfrak{a} = \langle \alpha_1, \alpha_2, \dots, \alpha_m \rangle$ and $\mathfrak{c} = \langle \gamma_1, \gamma_2, \dots, \gamma_p \rangle$. In that case \mathfrak{b} is generated by the mp products $\alpha_i \gamma_k$. However, each such product belongs to \mathfrak{a} (and also to \mathfrak{c} , but that's not important to us). It follows that every element of \mathfrak{b} is also an integer of \mathfrak{a} . ■

To summarize, it has been shown that every number system $\mathbb{Z}[\sqrt{d}]$ can be placed inside a new system of ideals in such a way that each element $\alpha \in \mathbb{Z}[\sqrt{d}]$ is placed on the ideal $\langle \alpha \rangle$. These ideals are subject to a binary operation, also called *ideal multiplication*, such that for any two integers $\alpha, \beta \in \mathbb{Z}[\sqrt{d}]$, $\langle \alpha \rangle \langle \beta \rangle = \langle \alpha\beta \rangle$.

In other words, the multiplication of ideals “faithfully represents” integer multiplication in the context of ideals and their multiplication. The advantage to algebraists is that the multiplication of ideals does possess unique prime factorization.

We conclude this section with two more examples for the reader to practice on:

Let $\mathfrak{p} = \langle 5, 2 + \sqrt{-6} \rangle$ and $\mathfrak{q} = \langle 2, \sqrt{-6} \rangle$. Standard calculations show that $\mathfrak{p}\bar{\mathfrak{p}} = \langle 5 \rangle$ and $\mathfrak{q}\bar{\mathfrak{q}} = \langle 2 \rangle$. Hence, $N(\mathfrak{p}) = 5$ and $N(\mathfrak{q}) = 2$, implying that

$$\langle 10 \rangle = \langle 5 \rangle \langle 2 \rangle = \mathfrak{p}\bar{\mathfrak{p}}\mathfrak{q}\bar{\mathfrak{q}}$$

is a factorization of the ideal $\langle 10 \rangle$ into prime ideals. The principal ideal $\langle 2 + \sqrt{-6} \rangle$ has norm 10.

Let $\mathfrak{p} = \langle 5, 2 + \sqrt{-6} \rangle$ and $\mathfrak{q} = \langle 7, 1 + \sqrt{-6} \rangle$. Then calculations show that

$$\mathfrak{p}\bar{\mathfrak{p}} = \langle 5 \rangle, \quad \mathfrak{q}\bar{\mathfrak{q}} = \langle 7 \rangle, \quad \mathfrak{p}\mathfrak{q} = \langle 35, -13 + \sqrt{-6} \rangle, \quad \text{and} \quad \bar{\mathfrak{p}}\bar{\mathfrak{q}} = \langle 35, -13 - \sqrt{-6} \rangle.$$

Hence the principal ideal $\langle 35 \rangle$ can be factored as

$$\langle 35 \rangle = \langle 5 \rangle \langle 7 \rangle = \mathfrak{p}\bar{\mathfrak{p}}\mathfrak{q}\bar{\mathfrak{q}}$$

as well as

$$\langle 35 \rangle = \langle 35, -13 + \sqrt{-6} \rangle \langle 35, -13 - \sqrt{-6} \rangle = \mathfrak{p}\mathfrak{q}\bar{\mathfrak{p}}\bar{\mathfrak{q}}.$$

Exercises 13.3

1. Decide whether the following assertions are true or false in $\mathbb{Z}[\sqrt{-6}]$.
 - (a) $\langle 2, 1 + \sqrt{-6} \rangle = \langle 2, 1 - \sqrt{-6} \rangle$
 - (b) $\langle 3, 1 + \sqrt{-6} \rangle = \langle 3, 1 - \sqrt{-6} \rangle$
 - (c) $\langle 2, 1 + \sqrt{-6} \rangle = \langle 4, 2 + 2\sqrt{-6} \rangle$
 - (d) $\langle 2, 1 + \sqrt{-6} \rangle = \langle 2, 1 + \sqrt{-6} \rangle$
 - (e) $\langle 29, 32 - 27\sqrt{-6} \rangle = \langle 3 + 2\sqrt{-6} \rangle$
 - (f) $\langle 49, 21 - 7\sqrt{-6}, 21 + 7\sqrt{-6}, 14 \rangle = \langle 7 \rangle$
 - (g) $\langle 3 - \sqrt{-6}, 1 + 2\sqrt{-6} \rangle = \langle 7, 3 - \sqrt{-6} \rangle$
2. Decide whether the following assertions are true or false in $\mathbb{Z}[\sqrt{-5}]$.
 - (a) $\langle 2, 1 + \sqrt{-5} \rangle = \langle 2, 1 - \sqrt{-5} \rangle$
 - (b) $\langle 3, 1 + \sqrt{-5} \rangle = \langle 3, 1 - \sqrt{-5} \rangle$
 - (c) $\langle 2, 1 + \sqrt{-5} \rangle = \langle 4, 2 + 2\sqrt{-5} \rangle$
 - (d) $\langle 2, 1 + \sqrt{-5} \rangle = \langle 2, 1 + \sqrt{-5} \rangle$
 - (e) $\langle 29, 32 - 27\sqrt{-5} \rangle = \langle 3 + 2\sqrt{-5} \rangle$
 - (f) $\langle 49, 21 - 7\sqrt{-5}, 21 + 7\sqrt{-5}, 14 \rangle = \langle 7 \rangle$
 - (g) $\langle 3 - \sqrt{-5}, 1 + 2\sqrt{-5} \rangle = \langle 7, 3 - \sqrt{-5} \rangle$
3. Which of these ideals are principal and which are not?

(a) $\langle 9 \rangle$	(f) $\langle 5, \sqrt{-6} \rangle$
(b) $\langle 2 - \sqrt{-6} \rangle$	(g) $\langle 2, 1 + \sqrt{-6} \rangle$
(c) $\langle 6, 8, 2 + 6\sqrt{-6} \rangle$	(h) $\langle 3, 1 + \sqrt{-6} \rangle$
(d) $\langle 3, 3\sqrt{-6} \rangle$	(i) $\langle 7, 1 + 2\sqrt{-6} \rangle$
(e) $\langle 3, \sqrt{-6} \rangle$	(j) $\langle 21, 9 + 3\sqrt{-6}, -2 + 4\sqrt{-6} \rangle$
4. Which of these ideals are principal and which are not?

(a) $\langle 9 \rangle$	(f) $\langle 5, \sqrt{-5} \rangle$
(b) $\langle 2 - \sqrt{-5} \rangle$	(g) $\langle 2, 1 + \sqrt{-5} \rangle$
(c) $\langle 6, 8, 2 + 6\sqrt{-5} \rangle$	(h) $\langle 3, 1 + \sqrt{-5} \rangle$
(d) $\langle 3, 3\sqrt{-5} \rangle$	(i) $\langle 7, 1 + 2\sqrt{-5} \rangle$
(e) $\langle 3, \sqrt{-5} \rangle$	(j) $\langle 21, 9 + 3\sqrt{-5}, -2 + 4\sqrt{-5} \rangle$
5. Prove that $\mathbb{Z}[-6]$ has no element of norm 2.

6. Complete the proof of Proposition 13.26.
7. Supply the missing details for the proof of Equation 13.28.
8. As proven in the text, in $\mathbb{Z}[\sqrt{-6}]$

$$\langle 2 + \sqrt{-6}, 7 + 2\sqrt{-6} \rangle = \langle 3, 1 - \sqrt{-6} \rangle.$$

Is this ideal $\langle 1 \rangle$?

13.4 Cancellation of Ideals

If $c \neq 0$ and a and b are all complex numbers, then it is well known that

$$ac = bc \Rightarrow a = b.$$

This property is not to be taken for granted. For example, in \mathbb{Z}_6

$$2 \cdot 3 \equiv 0 \equiv 4 \cdot 3 \quad \text{but} \quad 2 \not\equiv 4 \pmod{6}.$$

An ideal \mathfrak{c} of $\mathbb{Z}[\sqrt{d}]$ is a *cancelable ideal* if for every two ideals \mathfrak{a} and \mathfrak{b}

$$\mathfrak{a}\mathfrak{c} = \mathfrak{b}\mathfrak{c} \Rightarrow \mathfrak{a} = \mathfrak{b}.$$

It is clear that $\langle 0 \rangle$ is not cancelable since $\mathfrak{a}\langle 0 \rangle = \langle 0 \rangle = \mathfrak{b}\langle 0 \rangle$ for any ideals \mathfrak{a} and \mathfrak{b} .

Proposition 13.32 Nonzero principal ideals are cancelable.

Proof. Let $\mathfrak{a} = \langle \alpha_1, \alpha_2, \dots, \alpha_m \rangle$ and $\mathfrak{b} = \langle \beta_1, \beta_2, \dots, \beta_n \rangle$ be ideals and let $\langle \gamma \rangle$ be a nonzero principal ideal such that $\mathfrak{a}\langle \gamma \rangle = \mathfrak{b}\langle \gamma \rangle$ and consequently

$$\langle \gamma\alpha_1, \gamma\alpha_2, \dots, \gamma\alpha_m \rangle = \langle \gamma\beta_1, \gamma\beta_2, \dots, \gamma\beta_n \rangle.$$

This implies that every $\gamma\alpha_i$ is a $\mathbb{Z}[\sqrt{d}]$ -linear combination of the $\gamma\beta_j$'s, say

$$\gamma\alpha_i = \sum_{j=1}^n \alpha_{ij} \gamma\beta_j.$$

Cancellation by the nonzero integer γ displays α_i as a $\mathbb{Z}[\sqrt{d}]$ -linear combination of the β_j 's. Hence $\mathfrak{a} \subset \mathfrak{b}$. The reverse containment is proved in a similar manner. ■

Corollary 13.33 Every nonzero ideal which has a principal multiple is cancelable.

Proof. Let \mathfrak{c} be a nonzero ideal and suppose there exists another ideal \mathfrak{c}' such that $\mathfrak{c}\mathfrak{c}' = \langle \gamma \rangle$ for some nonzero integer γ . If \mathfrak{a} and \mathfrak{b} are any two ideals such that $\mathfrak{a}\mathfrak{c} = \mathfrak{b}\mathfrak{c}$ and hence

$$\mathfrak{a}\mathfrak{c}\mathfrak{c}' = \mathfrak{b}\mathfrak{c}\mathfrak{c}' \quad \text{or} \quad \mathfrak{a}\langle \gamma \rangle = \mathfrak{b}\langle \gamma \rangle.$$

Since $\langle \gamma \rangle$ is cancelable, it follows that $\mathfrak{a} = \mathfrak{b}$. ■

Lemma 13.34 If $\mathfrak{a} = \langle \alpha, \beta \rangle$ is an ideal of $\mathbb{Z}[\sqrt{d}]$ with two generators, then

$$\mathfrak{a}\bar{\mathfrak{a}} = \langle N(\alpha), \text{Tr}(\alpha\bar{\beta}), N(\beta) \rangle. \quad (13.35)$$

Proof. Clearly

$$\mathfrak{a}\bar{\mathfrak{a}} = \langle \alpha, \beta \rangle \langle \bar{\alpha}, \bar{\beta} \rangle = \langle \alpha\bar{\alpha}, \alpha\bar{\beta}, \beta\bar{\alpha}, \beta\bar{\beta} \rangle \supseteq \langle N(\alpha), \text{Tr}(\alpha\bar{\beta}), N(\beta) \rangle = \langle h \rangle$$

where, by Proposition 13.20, h is the greatest common divisor of the rational integers $N(\alpha)$, $\text{Tr}(\alpha\bar{\beta})$, and $N(\beta)$. This proves that the left-hand side of Equation 13.35 contains its right-hand side.

To prove the converse it suffices to show that $\alpha\bar{\beta}$ is contained in the right-hand side. This, in turn, will follow from proving that h divides $\alpha\bar{\beta}$ which is equivalent to saying that

$$\frac{\alpha\bar{\beta}}{h} \in \mathbb{Z}[\sqrt{d}].$$

This, however, follows from Proposition 13.3 and the equations

$$\text{Tr}\left(\frac{\alpha\bar{\beta}}{h}\right) = \frac{\alpha\bar{\beta} + \bar{\alpha}\beta}{h} = \frac{\text{Tr}(\alpha\bar{\beta})}{h} \in \mathbb{Z}[\sqrt{d}]$$

and

$$N\left(\frac{\alpha\bar{\beta}}{h}\right) = \frac{\alpha\bar{\beta}\bar{\alpha}\beta}{h^2} = \frac{N(\alpha)}{h} \frac{N(\beta)}{h} \in \mathbb{Z}[\sqrt{d}].$$
■

For example, let $\mathfrak{a} = \langle 2 + 3\sqrt{-6}, 4 \rangle$. Then $\mathfrak{a}\bar{\mathfrak{a}} = \langle 58, 8, 16 \rangle = \langle 2 \rangle$.

Theorem 13.36 If $\mathfrak{a} = \langle \alpha_1, \alpha_2, \dots, \alpha_m \rangle$, then $\mathfrak{a}\bar{\mathfrak{a}}$ is generated by the m rational integers $N(\alpha_1), N(\alpha_2), \dots, N(\alpha_m)$ and the $m(m-1)/2$ rational integers $\text{Tr}(\alpha_i\bar{\alpha}_j)$ where $i < j$.

Sketch of proof. Proceed by induction on m and use the method provided in the proof of Lemma 13.34. ■

Corollary 13.37 If \mathfrak{a} is an ideal, then $\mathfrak{a}\bar{\mathfrak{a}}$ is principal.

Proof. By Proposition 13.20, $\mathfrak{a}\bar{\mathfrak{a}}$ has a set of rational generators and $\mathfrak{a}\bar{\mathfrak{a}}$ is necessarily principal. ■

Corollary 13.38 Every nonzero ideal of $\mathbb{Z}[\sqrt{d}]$ is cancelable.

Proof. Let \mathfrak{a} be a nonzero ideal. By Corollary 13.37 $\mathfrak{a}\bar{\mathfrak{a}}$ is principal. In particular, \mathfrak{a} has a principal multiple. By Corollary 13.33, \mathfrak{a} is cancelable. ■

Corollary 13.39 If \mathfrak{a} and \mathfrak{b} are ideals of $\mathbb{Z}[\sqrt{d}]$, then

$$\mathfrak{a} \mid \mathfrak{b} \quad \text{if and only if} \quad \mathfrak{a} \supset \mathfrak{b}.$$

Proof. Suppose first that $\mathfrak{a} = \langle 0 \rangle$. Then the following are equivalent:

- $\mathfrak{a} \mid \mathfrak{b}$;
- $\langle 0 \rangle \mid \mathfrak{b}$;
- $\mathfrak{b} = \langle 0 \rangle$;
- $\langle 0 \rangle \supset \mathfrak{b}$.

Hence we may assume that $\mathfrak{a} \neq \langle 0 \rangle$. By Proposition 13.15, $\mathfrak{a} \mid \mathfrak{b}$ implies $\mathfrak{a} \supset \mathfrak{b}$. To prove the reverse implication, suppose $\mathfrak{a} \supset \mathfrak{b}$. It then follows that $\mathfrak{a}\bar{\mathfrak{a}} \supset \mathfrak{b}\bar{\mathfrak{a}}$. By Corollary 13.37 there exists an integer α such that $\mathfrak{a}\bar{\mathfrak{a}} = \langle \alpha \rangle$ and, hence, $\langle \alpha \rangle \supset \mathfrak{b}\bar{\mathfrak{a}}$, which implies that $\langle \alpha \rangle \mid \mathfrak{b}\bar{\mathfrak{a}}$.

By definition there exists a nonzero ideal \mathfrak{c} such that $\langle \alpha \rangle \mathfrak{c} = \mathfrak{b}\bar{\mathfrak{a}}$. When both sides of this equation are multiplied by \mathfrak{a} , we get

$$\langle \alpha \rangle \mathfrak{c}\mathfrak{a} = \mathfrak{b}\bar{\mathfrak{a}}\mathfrak{a} = \mathfrak{b}\langle \alpha \rangle.$$

Cancelation by the nonzero ideal $\mathfrak{a}\bar{\mathfrak{a}} = \langle \alpha \rangle$ yields $\mathfrak{c}\mathfrak{a} = \mathfrak{b}$ and hence $\mathfrak{a} \mid \mathfrak{b}$. ■

The gist of this surprising result is that amongst ideals “to contain is to divide.” This is somewhat counterintuitive because amongst the rational integers the divisor is generally viewed as being smaller than the dividend. In the context of ideals, however, the opposite happens: the divisor of an ideal is greater than the dividend in the sense that it contains the dividend.

Corollary 13.40 The divisors of an ideal \mathfrak{a} are precisely the ideals \mathfrak{b} such that $\mathfrak{b} \supset \mathfrak{a}$. In particular $\alpha \in \mathfrak{a}$ if and only if $\mathfrak{a} \mid \langle \alpha \rangle$.

This corollary makes it easy to find examples of proper containment (or division) amongst ideals. If $\mathfrak{a} = \langle \alpha_1, \alpha_2, \dots, \alpha_m \rangle$ and $\alpha \notin \mathfrak{a}$, then $\mathfrak{b} = \langle \alpha_1, \alpha_2, \dots, \alpha_m, \alpha \rangle$ is a proper divisor of \mathfrak{a} .

The *sum* of the two ideals \mathfrak{a} and \mathfrak{b} is

$$\mathfrak{a} + \mathfrak{b} = \{ \alpha + \beta \mid \alpha \in \mathfrak{a}, \beta \in \mathfrak{b} \}.$$

Proposition 13.41 If $\mathfrak{a} = \langle \alpha_1, \alpha_2, \dots, \alpha_m \rangle$ and $\mathfrak{b} = \langle \beta_1, \beta_2, \dots, \beta_n \rangle$, then

$$\mathfrak{a} + \mathfrak{b} = \langle \alpha_1, \alpha_2, \dots, \alpha_m, \beta_1, \beta_2, \dots, \beta_n \rangle$$

Proof. Exercise 13.4.2 ■

Proposition 13.42 For two ideals \mathfrak{a} and \mathfrak{b} , the ideal $\mathfrak{a} + \mathfrak{b}$ is a common divisor of \mathfrak{a} and \mathfrak{b} which is divisible by all the other common divisors.

Proof. It follows from Corollary 13.40 that $\mathfrak{a} + \mathfrak{b}$ divides both \mathfrak{a} and \mathfrak{b} . On the other hand, if \mathfrak{c} divides both \mathfrak{a} and \mathfrak{b} , then, by the same corollary, \mathfrak{c} contains both \mathfrak{a} and \mathfrak{b} so that $\mathfrak{c} \supset \mathfrak{a} + \mathfrak{b}$. This, of course, means that \mathfrak{c} divides $\mathfrak{a} + \mathfrak{b}$. ■

This proposition motivates the definition of the *greatest common divisor* of the two ideals \mathfrak{a} and \mathfrak{b} as their sum $\mathfrak{a} + \mathfrak{b}$.

For example, among the ideals of $\mathbb{Z}[\sqrt{-5}]$, the principal ideals $\langle 3 \rangle$ and $\langle 1 + \sqrt{-5} \rangle$ have greatest common divisor $\langle 3, 1 + \sqrt{-5} \rangle$, which is nonprincipal (Exercise 13.4.3).

Corollary 13.43 Every ideal is a sum of principal ideals.

Proof. Since every ideal in $\mathbb{Z}[\sqrt{d}]$ has the form $\langle \alpha_1, \alpha_2, \dots, \alpha_m \rangle$, it follows that

$$\begin{aligned} \langle \alpha_1, \alpha_2, \dots, \alpha_m \rangle &= \alpha_1 \mathbb{Z}[\sqrt{d}] + \alpha_2 \mathbb{Z}[\sqrt{d}] + \dots + \alpha_m \mathbb{Z}[\sqrt{d}] \\ &= \langle \alpha_1 \rangle + \langle \alpha_2 \rangle + \dots + \langle \alpha_m \rangle. \end{aligned}$$
■

Exercises 13.4

1. Find a single element of $a\bar{a}$ that generates it for the following ideals a :

(a) $\langle 12, 18 \rangle$

(c) $\langle 2 + 3\sqrt{-1}, 6 \rangle$

(b) $\langle 12, 18, 30 \rangle$

(d) $\langle 1 + 5\sqrt{-2}, 2 - \sqrt{-2} \rangle$

2. Prove Proposition 13.41.

3. Prove that $\langle 3, 1 + \sqrt{-6} \rangle$ is a nonprincipal ideal.

13.5 Norms of Ideals

According to Theorem 13.36, $a\bar{a}$ is a principal ideal that is generated by an element of \mathbb{Z} . Since for any such rational generator g , $\langle -g \rangle = \langle g \rangle$, we may assume that $g \geq 0$. The positive generator g is unique because if g and h are in \mathbb{Z} and $\langle g \rangle = \langle h \rangle$, then $g = hu$ for some unit U that is necessarily rational. Hence $u = \pm 1$.

If a is an ideal in $\mathbb{Z}[\sqrt{d}]$ let $N(a)$ denote the positive generator of $a\bar{a}$. We call $N(a)$ the (ideal) norm of a .

For example, in $\mathbb{Z}[\sqrt{-6}]$, the ideal $\langle 5, 2 + \sqrt{-6} \rangle$ has norm 5.

Proposition 13.44 The ideal norm and the element norm agree on principal ideals. That is, if $a = \langle \alpha \rangle$, then $N(a) = |N(\alpha)|$.

Proof. We have

$$a\bar{a} = \langle \alpha \rangle \langle \bar{\alpha} \rangle = \langle \alpha\bar{\alpha} \rangle = \langle N(\alpha) \rangle = \langle |N(\alpha)| \rangle.$$

The last term equals $\langle N(a) \rangle$, so $N(a) = |N(\alpha)|$ because both are positive and they generate the same principal ideal. ■

Proposition 13.45 If a and b are nonzero ideals, then $N(ab) = N(a)N(b)$.

Proof. We have

$$\langle N(ab) \rangle = ab\bar{a}\bar{b} = a\bar{a}b\bar{b} = a\bar{a}b\bar{b} = \langle N(a) \rangle \langle N(b) \rangle = \langle N(a)N(b) \rangle$$

from which it follows that $N(ab) = N(a)N(b)$ since both are positive generators of the same principal ideal. ■

Corollary 13.46 Let a and b be nonzero ideals such that $a \mid b$. Then $N(a) \mid N(b)$.

Proof. By hypothesis there exists an ideal c such that $b = ac$. By Proposition 13.45, $N(b) = N(a)N(c)$, which implies the desired conclusion. ■

The converse of the above corollary is false. The ideals $\mathfrak{a} = \langle 1 + \sqrt{-6} \rangle$ and $\mathfrak{b} = \langle 1 - \sqrt{-6} \rangle$ have the same norm of 15, yet neither divides the other.

In general, $N(\mathfrak{a}) = 1$ if and only if $\mathfrak{a} = \langle 1 \rangle$. In one direction this is trivial since clearly $N(\langle 1 \rangle) = 1$. Conversely, suppose $N(\mathfrak{a}) = 1$. Then $\mathfrak{a}\bar{\mathfrak{a}} = \langle 1 \rangle$, and so $\mathfrak{a} \mid 1$. This, however, implies that $\mathfrak{a} \supset \langle 1 \rangle = \mathbb{Z}[\sqrt{d}]$ and so $\mathfrak{a} = \langle 1 \rangle$.

One consequence of this observation is that the norm of every ideal is a positive rational integer. This will later facilitate proofs by mathematical induction on the norm.

It is okay to define $N(\langle 0 \rangle) = 0$, and all the results about ideal norms so far extend to zero ideals. That is not the case for the next property.

Corollary 13.47 If \mathfrak{a} is a nonzero ideal, then every ideal divisor of \mathfrak{a} , other than \mathfrak{a} , has norm less than $N(\mathfrak{a})$.

Proof. Let \mathfrak{b} be a factor of \mathfrak{a} other than \mathfrak{a} itself, so $\mathfrak{a} = \mathfrak{b}\mathfrak{c}$ and $\mathfrak{c} \neq \langle 1 \rangle$. Since $N(\mathfrak{a}) = N(\mathfrak{b}\mathfrak{c})$ with $N(\mathfrak{a}) \neq \langle 0 \rangle$ and $N(\mathfrak{c}) > 1$, it follows that $N(\mathfrak{b}) < N(\mathfrak{a})$. ■

Theorem 13.36 provides an algorithm for computing the norm of an arbitrary ideal. One need simply compute the greatest common divisor in \mathbb{Z} of all the listed norms and traces.

For example, let $\mathfrak{a} = \langle 5, 2 + \sqrt{-6} \rangle$. Then

$$\begin{aligned} \mathfrak{a}\bar{\mathfrak{a}} &= \langle 5 \cdot 5, 5(2 - \sqrt{-6}), 5(2 + \sqrt{-6}), (2 + \sqrt{-6})(2 - \sqrt{-6}) \rangle \\ &= \langle 5 \cdot 5, 5(2 - \sqrt{-6}), 5(2 + \sqrt{-6}), 10 \rangle \\ &= \langle 5 \rangle \langle 5, 2 - \sqrt{-6}, 2 + \sqrt{-6}, 2 \rangle \\ &= \langle 5 \rangle \langle 1 \rangle = \langle 5 \rangle \end{aligned}$$

since 5 and 2 are relatively prime rational integers. It follows that $N(\mathfrak{a}) = 5$.

Let $\mathfrak{a} = \langle 1 + \sqrt{-6}, 1 - \sqrt{-6} \rangle$. Since $\bar{\mathfrak{a}} = \mathfrak{a}$, it follows that

$$\begin{aligned} \mathfrak{a}\bar{\mathfrak{a}} &= \mathfrak{a}\mathfrak{a} = \langle (1 \pm \sqrt{-6})(1 \pm \sqrt{-6}) \rangle \\ &= \langle 1 - 6 + 2\sqrt{-6}, 1^2 - \sqrt{-6}^2, 1^2 - \sqrt{-6}^2, 1 - 6 - 2\sqrt{-6} \rangle \\ &= \langle -5 + 2\sqrt{-6}, 7, 7, -5 - 2\sqrt{-6} \rangle = \langle -5 + 2\sqrt{-6}, 7, -10 \rangle = \langle 1 \rangle. \end{aligned}$$

It follows that $N(\mathfrak{a}) = 1$.

Let $\mathfrak{a} = \langle 21 - 2\sqrt{-6}, 5 + \sqrt{-6} \rangle$. Note that $\mathfrak{a} = \langle 31, 5 + \sqrt{-6} \rangle$ so that

$$\begin{aligned} \mathfrak{a}\bar{\mathfrak{a}} &= \langle 31, 5 + \sqrt{-6} \rangle \langle 31, 5 - \sqrt{-6} \rangle \\ &= \langle 31^2, 31(5 + 2\sqrt{-6}), 31(5 - 2\sqrt{-6}), 5^2 + 1 \cdot 6 \rangle \\ &= \langle 31 \rangle \langle 31, (5 + 2\sqrt{-6}), 31(5 - 2\sqrt{-6}), 1 \rangle \\ &= \langle 31 \rangle. \end{aligned}$$

Hence, the norm of $\mathfrak{a} = 31$.

Exercises 13.5

1. Find the norms of the ideals of Exercise 13.4.1.

13.6 Prime Ideals and Unique Factorization

An ideal \mathfrak{p} is said to be a *prime ideal* if $\mathfrak{p} \neq \langle 1 \rangle$ and \mathfrak{p} cannot be written as a product $\mathfrak{a}\mathfrak{b}$ unless $\mathfrak{a} = \langle 1 \rangle$ or $\mathfrak{b} = \langle 1 \rangle$.

Note that in our treatment the zero ideal $\langle 0 \rangle$ is not considered to be a prime ideal. Here is a criterion for recognizing some prime ideals.

Proposition 13.48 An ideal whose norm is prime in \mathbb{Z} is a prime ideal.

Proof. Let $N(\mathfrak{a})$ be prime. If $\mathfrak{a} = \mathfrak{b}\mathfrak{c}$, then, by Proposition 13.45,

$$p = N(\mathfrak{b}\mathfrak{c}) = N(\mathfrak{b})N(\mathfrak{c}).$$

Since p is a rational prime, it follows that either $N(\mathfrak{a})$ or $N(\mathfrak{b})$ is 1. Hence either \mathfrak{a} or \mathfrak{b} is $\langle 1 \rangle$. ■

For example, as we saw earlier, in $\mathbb{Z}[\sqrt{-6}]$ the ideal $\langle 5, 2 + \sqrt{-6} \rangle$ has norm 5 and so this ideal is prime.

The converse to Proposition 13.48 is false: a prime ideal may have a composite norm. In $\mathbb{Z}[\sqrt{-22}]$, consider the principal ideal $\langle 17 \rangle$ whose norm is $289 = 17^2$. It will be shown that this ideal is in fact prime. Assume $\langle 17 \rangle = \mathfrak{a}\mathfrak{b}$ with $\mathfrak{a} \neq \langle 1 \rangle$ and $\mathfrak{b} \neq \langle 1 \rangle$. Taking norms we get $289 = N(\mathfrak{a})N(\mathfrak{b})$ and so $N(\mathfrak{a}) = 17$. Suppose $\mathfrak{a} = \langle \alpha_1, \alpha_2, \dots, \alpha_m \rangle$. Clearly $\mathfrak{a} \mid \langle \alpha_i \rangle$ for each $i = 1, 2, \dots, m$ and so $17 \mid N(\alpha_i)$. We now show that, in fact, $17 \mid \alpha_i$ for all $i = 1, 2, \dots, m$. Suppose $\alpha_i = x + y\sqrt{-22}$. Since we are stipulating that

$17 \mid N(\alpha_i)$ it follows that

$$x^2 + 22y^2 \equiv 0 \pmod{17} \quad \text{or} \quad x^2 \equiv -5y^2 \pmod{17}.$$

Since -5 is not a quadratic residue modulo 17, it must be the case that $y \equiv 0 \pmod{p}$ and hence also $x \equiv 0 \pmod{p}$. This, of course, implies that *every* such integer of \mathfrak{a} is divisible by 17. Let \mathfrak{c} be the ideal obtained from \mathfrak{a} by dividing each integer of \mathfrak{a} by 17. Then $\mathfrak{a} = \langle 17 \rangle \mathfrak{c}$. Applying norms to both sides yields

$$N(\mathfrak{a}) = N(\langle 17 \rangle)N(\mathfrak{c}) \geq 17^2 \cdot 1 = 289$$

which contradicts the fact that $N(\mathfrak{a}) = 17$. Hence $\langle 17 \rangle$ is a prime ideal of $\mathbb{Z}[\sqrt{-22}]$ whose norm is the composite number 289.

Proposition 13.49 If an ideal is prime, then its conjugate ideal is also prime.

Proof. See Exercise 13.6.6. ■

We now set out to prove that just like the rational integers, the ideals of $\mathbb{Z}[\sqrt{d}]$ have the unique factorization property. The proof for the irrational quadratic case is very similar to that for the rational integers. Recall that the proof that the rational integers have the unique factorization property has three steps:

- (a) show that prime numbers satisfy the property $p \mid ab \Rightarrow p \mid a$ or $p \mid b$;
- (b) show by induction that every positive integer greater than 1 has a prime factorization; and
- (c) show by induction that the prime factorization is unique, using the first step and cancellation to reduce to a smaller case.

The following proposition is an analog of the first step.

Proposition 13.50 If \mathfrak{p} is a prime ideal and $\mathfrak{p} \mid \mathfrak{a}\mathfrak{b}$, then $\mathfrak{p} \mid \mathfrak{a}$ or $\mathfrak{p} \mid \mathfrak{b}$.

Proof. We will assume that \mathfrak{p} divides $\mathfrak{a}\mathfrak{b}$ but does not divide \mathfrak{a} . The ideal $\mathfrak{p} + \mathfrak{a}$ is a common divisor of \mathfrak{p} and \mathfrak{a} since it contains both. Since \mathfrak{p} is a prime ideal, the only divisors of \mathfrak{p} are \mathfrak{p} and $\langle 1 \rangle$. Because it was assumed that \mathfrak{p} does not divide \mathfrak{a} , $\mathfrak{p} + \mathfrak{a} \neq \mathfrak{p}$, since otherwise $\mathfrak{p} + \mathfrak{a} = \mathfrak{p}$ would imply that $\mathfrak{a} \subset \mathfrak{p}$ meaning that \mathfrak{p} contains \mathfrak{a} , or \mathfrak{p} divides \mathfrak{a} , contradicting the assumption that begins this proof. Therefore $\mathfrak{p} + \mathfrak{a} = \langle 1 \rangle$, and hence $1 = \pi + \alpha$ for some $\pi \in \mathfrak{p}$ and $\alpha \in \mathfrak{a}$. Hence for any $\beta \in \mathfrak{b}$

$$\beta = 1 \cdot \beta = \pi\beta + \alpha\beta \in \mathfrak{p} + \mathfrak{a}\mathfrak{b} = \mathfrak{p},$$

showing that $b \in p$. Thus, $p \mid b$. ■

Corollary 13.51 If p is prime and $p \mid a_1 a_2 \cdots a_r$, then $p \mid a_i$ for some i .

Proof. By induction on r (see Exercise 13.6.7). ■

We now work out the analog of the second step toward prime factorizations.

Proposition 13.52 Every nonzero ideal $\neq \langle 1 \rangle$ admits a prime ideal factorization.

Proof. Mimic the proof of Theorem 4.9 by using induction on the ideal norm. ■

Proposition 13.53 The prime factorization of any nonzero ideal $\neq \langle 1 \rangle$ is unique up to the order of the factors. That is, for any nonzero $a \neq \langle 1 \rangle$, if

$$a = p_1 p_2 \cdots p_r = q_1 q_2 \cdots q_s$$

where the p_i 's and q_j 's are prime ideals, then $r = s$ and

$$p_i = q_i \quad \text{for all } i = 1, 2, \dots, r$$

after a suitable reindexing.

Proof. By induction on the norm of the ideal. A prime ideal, by definition, has a unique prime factorization, namely, itself. This verifies the proposition for all prime ideals, including those of norm 2. For norm $n \geq 3$, suppose that the proposition holds for all ideals of norm $2, 3, \dots, n-1$. Let a be an ideal of norm n with two prime factorizations

$$a = p_1 p_2 \cdots p_r = q_1 q_2 \cdots q_s. \quad (13.54)$$

Since this ideal is not prime, we may assume that $r > 1$ and $s > 1$. Because $p_1 \mid a$, it follows from Corollary 13.51 above that $p_1 \mid q_j$ for some $j = 1, 2, \dots, s$. The commutativity of ideal multiplication (Proposition 13.17) makes it possible to relabel the ideals q_j , if necessary, so that $p_1 \mid q_1$. Since all nonzero ideals are cancelable (Corollary 13.38), it follows that

$$p_2 p_3 \cdots p_r = q_2 q_3 \cdots q_s.$$

As the left-hand side has a norm that is smaller than n , the induction hypothesis is applicable. Hence $r-1 = s-1$, or $r = s$, and the ideals of the right-hand side can be relabeled so that $p_i = q_i$ for all $i = 1, 2, \dots, r$. ■

Theorem 13.55 The integers of $\mathbb{Z}[\sqrt{d}]$ have the unique factorization property if and only if every ideal of $\mathbb{Z}[\sqrt{d}]$ is principal.

Proof. First, suppose the integers of $\mathbb{Z}[\sqrt{d}]$ have the unique factorization property.

Step 1: For every irreducible $\pi \in \mathbb{Z}[\sqrt{d}]$, the principal ideal $\langle \pi \rangle$ is prime. Let \mathfrak{a} be an ideal that divides $\langle \pi \rangle$, so that $\mathfrak{a} \supset \langle \pi \rangle$. It suffices to show that \mathfrak{a} is either $\langle 1 \rangle$ or $\langle \pi \rangle$. Suppose $\mathfrak{a} \neq \langle \pi \rangle$ so that there is an element $\alpha \in \mathfrak{a}$ such that $\alpha \notin \langle \pi \rangle$. By definition, there exists an ideal \mathfrak{b} such that $\pi = \alpha\mathfrak{b}$. Clearly $\mathfrak{b} \mid \pi$ and for every $\beta \in \mathfrak{b}$ $\alpha\beta \in \alpha\mathfrak{b} = \langle \pi \rangle$, and so $\pi \mid \alpha\beta$.

By the unique factorization of integers π is an irreducible factor of either α or β . Since π does not divide α , it must be that $\pi \mid \beta$, and so $\mathfrak{b} \subset \langle \pi \rangle$ since β is an arbitrary element of \mathfrak{b} .

Step 2: Every prime ideal in $\mathbb{Z}[\sqrt{d}]$ is principal. Let \mathfrak{p} be a prime ideal. Then $\mathfrak{p} \mid \langle a \rangle$ for some nonzero rational integer a (one can always use $a = N(\mathfrak{p})$). Factor a into irreducibles in $\mathbb{Z}[\sqrt{d}]$, say $a = \pi_1\pi_2\cdots\pi_r$. Then $\langle a \rangle = \langle \pi_1 \rangle \langle \pi_2 \rangle \cdots \langle \pi_r \rangle$ and so \mathfrak{p} divides some $\langle \pi_i \rangle$. Since this $\langle \pi_i \rangle$ is prime by Step 1, it follows that $\mathfrak{p} = \langle \pi_i \rangle$.

Step 3: Every ideal in $\mathbb{Z}[\sqrt{d}]$ is principal. The zero ideal is clearly principal. Every nonzero ideal is a product of prime ideals which are principal by Step 2. Since the product of principal ideals is principal (Proposition 13.22), \mathfrak{p} is also principal. This concludes the only if (\Rightarrow) part of the proof.

To prove the converse assume that every ideal in $\mathbb{Z}[\sqrt{d}]$ is principal. The existence of a factorization into irreducibles is guaranteed by Proposition 13.50. To prove uniqueness it suffices to show that for any irreducible π , $\pi \mid \alpha\beta$ implies that either $\pi \mid \alpha$ or $\pi \mid \beta$. (Recall that the analog of this property for prime ideals in the proof of Proposition 13.50 was used to prove the uniqueness of prime factorizations.)

So suppose $\pi \mid \alpha\beta$ and π does not divide α . The only factors of π are units and unit multiples of π , so the only common divisors of π and α are units. The ideal $\langle \pi, \alpha \rangle$ is principal by hypothesis, say $\langle \pi, \alpha \rangle = \langle \delta \rangle$. It follows that δ is a unit, so $\langle \pi, \alpha \rangle = \langle 1 \rangle$. Thus, there exist integers $x, y \in \mathbb{Z}[\sqrt{d}]$ such that $\pi x + \alpha y = 1$. Multiplying through by β we get

$$\pi\beta x + \alpha\beta y = \beta.$$

Since $\pi \mid \alpha\beta$, it follows that $\pi \mid \beta$. ■

Exercises 13.6

1. Explain why units do not appear in either the statement or the proof of Proposition 13.53.
2. Is the ideal $\langle 5 \rangle$ prime in $\mathbb{Z}[-5]$?
3. Is the ideal $\langle 5 \rangle$ prime in $\mathbb{Z}[-6]$?
4. Is the ideal $\langle 6 \rangle$ prime in $\mathbb{Z}[-5]$?
5. Is the ideal $\langle 6 \rangle$ prime in $\mathbb{Z}[-6]$?
6. Prove Proposition 13.49.
7. Prove Corollary 13.51.

13.7 Constructing Prime Ideals

The prime ideals of $\mathbb{Z}[\sqrt{d}]$ will now be classified in a scheme similar to that used for the Gaussian integers (Theorem 12.48).

Theorem 13.56 Every prime ideal in $\mathbb{Z}[\sqrt{d}]$ divides a unique prime rational integer. That is, if \mathfrak{p} is prime, then $\mathfrak{p} \mid \langle p \rangle$ for exactly one prime p of \mathbb{Z}^+ .

Proof. The ideal $\mathfrak{p}\bar{\mathfrak{p}} = \langle N(\mathfrak{p}) \rangle$ is divisible by \mathfrak{p} and has a generator in \mathbb{Z}^+ . Since $\mathfrak{p} \neq \langle 1 \rangle$, $N(\mathfrak{p}) > 1$.

Factor $N(\mathfrak{p})$ into primes in \mathbb{Z}^+ , say

$$N(\mathfrak{p}) = p_1 p_2 \cdots p_r.$$

Then

$$\mathfrak{p}\bar{\mathfrak{p}} = \langle p_1 p_2 \cdots p_r \rangle = \langle p_1 \rangle \langle p_2 \rangle \cdots \langle p_r \rangle,$$

so \mathfrak{p} divides some α_i by Corollary 13.51.

For the uniqueness, assume that $\mathfrak{p} \mid \langle p \rangle$ and $\mathfrak{p} \mid \langle q \rangle$ for two different prime rational integers. Since p and q are relatively prime, it follows from Proposition 13.20 that $\mathfrak{p} = \langle 1 \rangle$, a contradiction. ■

Corollary 13.57 Every prime ideal of $\mathbb{Z}[\sqrt{d}]$ has norm p or p^2 for some rational prime p .

Proof. Let \mathfrak{p} be a prime ideal in $\mathbb{Z}[\sqrt{d}]$. Then there is a rational prime integer p such that $\mathfrak{p} \mid \langle p \rangle$. Taking ideal norms, we get $N(\mathfrak{p}) \mid N(\langle p \rangle)$. Since

$$N(\langle p \rangle) = |N(\langle p \rangle)| = p^2,$$

it follows that $N(\mathfrak{p})$ is p or p^2 . ■

The significance of Theorem 13.56 is that it facilitates the discovery of all the prime ideals of $\mathbb{Z}[\sqrt{d}]$.

The following theorem describes how each prime number (really the principal ideal generated by each prime number) factors in $\mathbb{Z}[\sqrt{d}]$ and thus shows us what all the prime ideals of $\mathbb{Z}[\sqrt{d}]$ look like.

Theorem 13.58 For each (rational) prime number p :

- (a) If d is a quadratic nonresidue of p , then $\langle p \rangle$ is prime in $\mathbb{Z}[\sqrt{d}]$ with norm p^2 .
- (b) If d is a quadratic residue modulo p with distinct square roots c and c' , then $\langle p \rangle = \mathfrak{p}\bar{\mathfrak{p}}$ where \mathfrak{p} has norm p , and $\mathfrak{p} = \langle p, \sqrt{d} - c \rangle$.
- (c) If d is a quadratic residue modulo p with a double square root c , then $\langle p \rangle = \mathfrak{p}^2$ where \mathfrak{p} has norm p , and $\mathfrak{p} = \langle p, \sqrt{d} - c \rangle$.

Proof. Proof of (a): We show that if p is a rational prime and d is a quadratic nonresidue of p , then $\langle p \rangle$ is a prime ideal. Suppose not. Then there exists a nontrivial factorization $\langle p \rangle = \mathfrak{a}\mathfrak{b}$. A standard norm argument leads to the conclusion that $N(\mathfrak{a}) = p$. Thus, \mathfrak{p} is a prime ideal. Since that $\langle p \rangle \subset \mathfrak{a}$ and $\langle p \rangle \neq \mathfrak{a}$, it follows that there exists an element $a + b\sqrt{d}$ such that

$$a + b\sqrt{d} \in \mathfrak{a}, \quad a + b\sqrt{d} \notin \langle p \rangle.$$

Since $\mathfrak{a} \mid \langle a + b\sqrt{d} \rangle$, it follows by taking norms that $p \mid N(a + b\sqrt{d})$, or

$$a^2 - db^2 \equiv 0 \pmod{p}.$$

If $b \equiv 0 \pmod{p}$, then necessarily also $a \equiv 0 \pmod{p}$, which contradicts the fact that $a + b\sqrt{d}$ is not in $\langle p \rangle$. Hence $b \not\equiv 0 \pmod{p}$. It follows that when the equation above is divided by b we get

$$(ab^{-1})^2 \equiv d \pmod{p},$$

contradicting the assumption that d is a quadratic nonresidue modulo p . In other words if d is a quadratic nonresidue modulo p , then $\langle p \rangle$ is a prime ideal.

Proof of (b): Next suppose p is a rational prime and d is a quadratic residue modulo p with distinct roots c and c' . Set

$$\mathfrak{p} = \langle p, \sqrt{d} - c \rangle.$$

Since $p \in \mathfrak{p}$, $\mathfrak{p} \mid \langle p \rangle$. Trivially, $\sqrt{d} - c \in \mathfrak{p}$. By Proposition 13.4, $\sqrt{d} - c \notin \langle p \rangle$ and hence $\mathfrak{p} \neq \langle p \rangle$. Therefore, $N(\mathfrak{p})$ properly divides p^2 and so $N(\mathfrak{p})$ is 1 or p . Clearly,

$$\bar{\mathfrak{p}} = \langle p, -\sqrt{d} - c \rangle$$

and hence

$$\begin{aligned} \mathfrak{p}\bar{\mathfrak{p}} &= \langle p, \sqrt{d} - c \rangle \langle p, -\sqrt{d} - c \rangle \\ &= \langle p^2, p(\sqrt{d} - c), p(-\sqrt{d} - c), c^2 - d \rangle. \end{aligned}$$

Since $c^2 - d \equiv 0 \pmod{p}$, it follows that

$$\mathfrak{p}\bar{\mathfrak{p}} = \langle p \rangle \langle p, \sqrt{d} - c, -\sqrt{d} - c, (c^2 - d)/p \rangle.$$

Since both the ideals $\mathfrak{p}\bar{\mathfrak{p}}$ and $\langle p \rangle$ have norm p^2 it follows that the norm of $\langle p, \sqrt{d} - c, -\sqrt{d} - c, (c^2 - d)/p \rangle$ is 1 and hence this is the unit ideal so that $\mathfrak{p}\bar{\mathfrak{p}} = \langle p \rangle$.

Note that $N(\mathfrak{p}) = p$, where p is prime, implies that \mathfrak{p} is a prime ideal. It remains to show that $\mathfrak{p} \neq \bar{\mathfrak{p}}$. Assume otherwise, so that

$$\begin{aligned} \langle p \rangle = \mathfrak{p}^2 &= \langle p, \sqrt{d} - c \rangle \langle p, \sqrt{d} - c \rangle \\ &= \langle p^2, p(\sqrt{d} - c), p(\sqrt{d} - c), (\sqrt{d} - c)^2 \rangle \\ &= \langle p^2, p(\sqrt{d} - c), (\sqrt{d} - c)^2 \rangle \\ &= \langle p^2, p(\sqrt{d} - c), \sqrt{d}^2 - 2c\sqrt{d} + c^2 \rangle. \end{aligned}$$

In order for this to equal $\langle p \rangle$ the third generator needs to be divisible by p . However, we already know that

$$c^2 + \sqrt{d}^2 = c^2 + d \equiv 0 \pmod{p}.$$

Hence $p \mid 2c\sqrt{d}$. Similarly, $p \mid 2c'\sqrt{d}$ and hence it follows that

$$c \equiv c' \pmod{p},$$

which contradicts the assumption that c and c' are distinct modulo p .

Proof of (c): In the remaining case d is a quadratic residue modulo p with a double square root c .

Set $\mathfrak{p} = (p, \sqrt{d})$. Arguing as was done above, $N(\mathfrak{p})$ is p and hence \mathfrak{p} is a prime ideal. Next we compute

$$\begin{aligned} \mathfrak{p}^2 &= \langle p, \sqrt{d} - c \rangle \langle p, \sqrt{d} - c \rangle \\ &= \langle p^2, p(\sqrt{d} - c), (\sqrt{d} - c)^2 \rangle \\ &= \langle p^2, p(\sqrt{d} - c), d - 2c\sqrt{d} + c^2 \rangle, \end{aligned}$$

since $\sqrt{d}^2 = d$.

However, $c^2 + d \equiv 0 \pmod{p}$ and $c \equiv 0 \pmod{p}$ and hence both $c^2 + d$ and $2c\sqrt{d}$ are divisible by p . Consequently $\langle p \rangle$ can be factored from the four generators of \mathfrak{p}^2 above, resulting in

$$\mathfrak{p}^2 = \langle p \rangle \langle p, (\sqrt{d} - c), (-2c\sqrt{d} + c^2 + d)/p \rangle.$$

Taking norms on both sides shows that the last ideal on the right has norm 1 and hence is $\langle 1 \rangle$. ■

We demonstrate several applications of this theorem:

Is $\langle 3 \rangle$ a prime ideal in $\mathbb{Z}[\sqrt{-1}]$? Since -1 is a quadratic nonresidue modulo 3, it follows from Theorem 13.58 that $\langle 3 \rangle$ is a prime ideal in $\mathbb{Z}[\sqrt{-1}]$.

Is $\langle 5 \rangle$ a prime ideal in $\mathbb{Z}[\sqrt{-1}]$? Since -1 is a quadratic residue modulo 5, in fact $-1 \equiv (\pm 2)^2$, and we chose the value 2 for c , then

$$\langle 5 \rangle = \langle 5, -2 + \sqrt{-1} \rangle \langle 5, -2 - \sqrt{-1} \rangle$$

is the prime factorization of the ideal $\langle 5 \rangle$ (see Exercise 13.7.6).

Is $\langle 5 \rangle$ a prime ideal in $\mathbb{Z}[\sqrt{-6}]$? Since

$$-6 \equiv 2^2 \equiv 3^2 \pmod{5}$$

the ideal $\langle 5 \rangle$ has the prime ideal factorization

$$\langle 5, \sqrt{-6} - 2 \rangle \langle 5, -\sqrt{-6} - 2 \rangle.$$

Is $\langle 7 \rangle$ a prime ideal in $\mathbb{Z}[\sqrt{-6}]$? Here $1^2 \equiv 6^2 \equiv 1 \pmod{7}$ and hence $\langle 7 \rangle$ factors into

$$\langle 7, \sqrt{-6} - 1 \rangle \langle 7, -\sqrt{-6} - 1 \rangle.$$

Is $\langle 11 \rangle$ a prime ideal in $\mathbb{Z}[\sqrt{-14}]$? As -14 is a quadratic nonresidue modulo 11, it follows that $\langle 11 \rangle$ is a prime ideal of $\mathbb{Z}[\sqrt{-14}]$.

Is $\langle 7 \rangle$ a prime ideal in $\mathbb{Z}[\sqrt{-14}]$? Zero is a quadratic nonresidue of any prime modulus. Consequently $\langle 7 \rangle$ is a prime.

Is $\langle 2 \rangle$ a prime ideal in $\mathbb{Z}[\sqrt{-1}]$? Since -1 is a quadratic square modulo 2 with a double square root, it follows from part 3 of Theorem 13.58 that

$$\langle 2 \rangle = \langle 1 + \sqrt{-1} \rangle^2$$

where $\langle 1 + \sqrt{-1} \rangle$ is a prime ideal of $\mathbb{Z}[\sqrt{-1}]$.

Listed in Tables 13.1 and 13.2 are summaries of the calculations leading to “small” prime ideals of $\mathbb{Z}[\sqrt{d}]$ for $d = -5$ and -6 . The entry in the column marked p can be any rational prime integer.

For example, to factor the principal ideal $\mathfrak{a} = \langle 2 + \sqrt{-6} \rangle$ into prime ideals, notice first that the norm of $\langle 2 + \sqrt{-6} \rangle$ is $2^2 - (-6) = 10$. Hence the norms of the prime factors of \mathfrak{a} are 2 and 5, and consequently $2 + \sqrt{-6}$ has two factors: one from the pair $\{\mathfrak{p}_2, \bar{\mathfrak{p}}_2\}$ and one from the pair $\{\mathfrak{p}_5, \bar{\mathfrak{p}}_5\}$. However, since $\{ \mathfrak{p}_2 = \bar{\mathfrak{p}}_2 \}$, it follows that \mathfrak{p}_2 is a prime factor of $2 + \sqrt{-6}$. On the other hand, $\mathfrak{p}_5 \in \langle 2 + \sqrt{-6}, 5 \rangle$ and hence \mathfrak{p}_5 is the other prime factor of $\langle 2 + \sqrt{-6} \rangle$. Thus

$$\langle 2 + \sqrt{-6} \rangle = \mathfrak{p}_2 \mathfrak{p}_5$$

is the required prime factorization.

p	$-5 \equiv c^2 \pmod{p}$	\mathfrak{p}_p	$\langle p \rangle$	$N(\mathfrak{p}_p)$
2	$-5 \equiv 1 \equiv c^2 \pmod{2}$	$\langle 1 + \sqrt{-5}, 2 \rangle$	$\mathfrak{p}_2 \bar{\mathfrak{p}}_2$	2
3	$-5 \equiv 1 \equiv c^2 \pmod{3}$	$\langle 1 + \sqrt{-5}, 3 \rangle$	$\mathfrak{p}_3 \bar{\mathfrak{p}}_3$	3
5	$-5 \equiv 0 \equiv c^2 \pmod{5}$	$\langle \sqrt{-5}, 5 \rangle = \langle \sqrt{-5} \rangle$	\mathfrak{p}_5^2	5
7	$-5 \equiv 2 \equiv c^2 \pmod{7}$	$\langle 3 + \sqrt{-5}, 7 \rangle$	$\mathfrak{p}_7 \bar{\mathfrak{p}}_7$	7
11	$-5 \equiv 6 \equiv c^2 \pmod{11}$	$\langle 11 \rangle$	$\langle 11 \rangle$	11^2
13	$-5 \equiv 8 \equiv c^2 \pmod{13}$	$\langle 13 \rangle$	$\langle 13 \rangle$	13^2
17	$-5 \equiv 12 \equiv c^2 \pmod{17}$	$\langle 17 \rangle$	$\langle 17 \rangle$	17^2
19	$-5 \equiv 14 \equiv c^2 \pmod{19}$	$\langle 19 \rangle$	$\langle 19 \rangle$	19^2
23	$-5 \equiv 18 \equiv c^2 \pmod{23}$	$\langle 23 \rangle$	$\langle 23 \rangle$	23^2

Table 13.1 Some calculations in $\mathbb{Z}[\sqrt{d}]$

p	$-6 \equiv c^2 \pmod{p}$	\mathfrak{p}_p	$\langle p \rangle$	$N(\mathfrak{p}_p)$
2	$-6 \equiv 0 \equiv x^2 \pmod{2}$	$\langle \sqrt{-6}, 2 \rangle$	\mathfrak{p}_2^2	2
3	$-6 \equiv 0 \equiv x^2 \pmod{3}$	$\langle \sqrt{-6}, 3 \rangle$	\mathfrak{p}_3^2	3
5	$-6 \equiv 4 \equiv x^2 \pmod{5}$	$\langle 2 + \sqrt{-6}, 5 \rangle$	$\mathfrak{p}_5 \bar{\mathfrak{p}}_5$	5
7	$-6 \equiv 1 \equiv x^2 \pmod{7}$	$\langle 1 + \sqrt{-6}, 7 \rangle$	$\mathfrak{p}_7 \bar{\mathfrak{p}}_7$	7
11	$-6 \equiv 5 \equiv x^2 \pmod{11}$	$\langle 4 + \sqrt{-6}, 11 \rangle$	$\mathfrak{p}_{11} \bar{\mathfrak{p}}_{11}$	11
13	$-6 \equiv 7 \equiv x^2 \pmod{13}$	$\langle 13 \rangle$	$\langle 13 \rangle$	13^2
17	$-6 \equiv 11 \equiv x^2 \pmod{17}$	$\langle 17 \rangle$	$\langle 17 \rangle$	17^2
19	$-6 \equiv 13 \equiv x^2 \pmod{19}$	$\langle 19 \rangle$	$\langle 19 \rangle$	19^2
23	$-6 \equiv 17 \equiv x^2 \pmod{23}$	$\langle 23 \rangle$	$\langle 23 \rangle$	23^2

Table 13.2 More calculations in $\mathbb{Z}[\sqrt{d}]$

To factor the principal ideal $\langle 4 + 3\sqrt{-6} \rangle$ into prime ideals, the norm of $4 + 3\sqrt{-6}$ is $4^2 - 3^2(-6) = 2 \cdot 5 \cdot 7$. Hence the norms of the prime factors of $4 + 3\sqrt{-6}$ have norms 2, 5, and 7 and consequently this ideal has as prime factors \mathfrak{p}_2 , one of the pair $\{\mathfrak{p}_5, \bar{\mathfrak{p}}_5\}$, and one of the pair $\{\mathfrak{p}_7, \bar{\mathfrak{p}}_7\}$.

The ideal \mathfrak{p}_5 is a factor of $4 + 3\sqrt{-6}$ if and only if \mathfrak{p}_5 divides $\langle 4 + 3\sqrt{-6} \rangle$ and this happens if and only if

$$4 + 3\sqrt{-6} \in \langle 2 + \sqrt{-6}, 5 \rangle.$$

This is done by the method of example; namely, we look for integers $\alpha = a + b\sqrt{-6}$ and $\beta = c + d\sqrt{-6}$ such that $4 + 3\sqrt{-6} = \alpha(2 + \sqrt{-6}) + \beta(5)$, or

$$4 + 3\sqrt{-6} = (a + b\sqrt{-6})(2 + \sqrt{-6}) + (c + d\sqrt{-6})(5).$$

Equating real and purely imaginary parts we get $4 = 2a - 6b + 5c$ and $3 = a + 2b + 5d$. One way to find integer solutions of these two equations is to rewrite them as $5c = 4 - 2a + 6b$ and $5d = -a - 2b - 3$, which have the solution $a = 0$, $b = 1$, $c = 2$, $d = -1$. The existence of this solution guarantees that \mathfrak{p}_5 is a prime ideal of $\langle 4 + 3\sqrt{-6} \rangle$.

Finally, we turn to \mathfrak{p}_7 . It divides $\langle 4 + 3\sqrt{-6} \rangle$ if and only if $4 + 3\sqrt{-6} \in \langle 1 + \sqrt{-6}, 7 \rangle$. The same substitution that was used above leads to the system of equations $4 = a - 6b + 7c$ and $3 = a + b + 7d$. When these two equations are subtracted from each other, we obtain $1 = -7b + 7c - 7d$, which contradicts the requirement that b , c , and d are rational integers. Hence, \mathfrak{p}_7 does not divide $\langle 1 + \sqrt{-6}, 7 \rangle$, so that $\bar{\mathfrak{p}}_7$ does divide it. Consequently

$$\langle 4 + 3\sqrt{-6} \rangle = \mathfrak{p}_2 \mathfrak{p}_5 \bar{\mathfrak{p}}_7.$$

Exercises 13.7

1. Construct the analog of Table 13.1 for $d = -10$.
2. Construct the analog of Table 13.1 for $d = -13$.
3. Construct the analog of Table 13.1 for $d = -14$.
4. Prove that every ideal divides some rational integer.
5. Let \mathcal{J} be an ideal and suppose α and β are rational primes in \mathcal{J} . Prove that α and β are associates of each other.
6. Prove that $\langle 5 \rangle = \langle 5, -2 + \sqrt{-1} \rangle \langle 5, -2 - \sqrt{-1} \rangle$ in $\mathbb{Z}[\sqrt{-1}]$.

Chapter Summary

We constructed a large number of domains whose ideals do provide unique factorizations into primes.

Chapter Review Exercises

Mark the following true or false.

1. Every nonzero nonunit in $\mathbb{Z}[\sqrt{d}]$ is a product of irreducibles in $\mathbb{Z}[\sqrt{d}]$.
2. Ideals are real.
3. The only nonprincipal ideal of $\mathbb{Z}[\sqrt{-5}]$ is $\langle 3, 1 + \sqrt{-5} \rangle$.

New Terms

associates, 320	norm, 318
cancelable ideal, 337	prime ideal, 343
conjugate, 318	principal ideal, 323
conjugate ideal, 331	product, 326
greatest common divisor, 340	quadratic field, 318
ideal multiplication, 335	trace, 318
ideals, 322	unit, 320
irreducible, 320	

Chapter 14



ABSTRACT RINGS

THE NOTION of a *ring* was introduced by David Hilbert circa 1896 in order to provide axiomatic descriptions of such mathematical structures as \mathbb{R} , \mathbb{Q} , \mathbb{Z} , \mathbb{Z}_n , $\mathbb{Z}[\sqrt{d}]$, $F[X]$, and many others. They all possess two mathematical operations that behave very much like the standard, school-taught arithmetic. It therefore made sense to isolate them and identify them by an appropriate nomenclature.

14.1 Rings

A (*commutative*) *ring (with unity)* is a set R with two binary operations, usually denoted by $+$ and \cdot , and two distinct special elements, usually denoted by 0 and 1 , for which the following hold: For any elements a, b, c of R

$a + b \in R$	$a \cdot b \in R$	(closure)
$(a + b) + c = a + (b + c)$	$(a \cdot b) \cdot c = a \cdot (b \cdot c)$	(associativity)
$a + b = b + a$	$a \cdot b = b \cdot a$	(commutativity)
$a \cdot (b + c) = a \cdot b + a \cdot c$		(distributivity)

there exist distinct elements 0 and 1 in R such that

$$a + 0 = a \quad a \cdot 1 = a \quad (\text{identities})$$

and there exists an element $-a \in F$ such that

$$a + (-a) = 0 \quad (\text{additive inverse}).$$

Note that these properties are nearly identical with those of a field, except for the issue of the existence of multiplicative inverses. The nonzero elements of rings are not required

to possess multiplicative inverses. Consequently, every field is a ring but rings need not be fields. The integers \mathbb{Z} do constitute a ring, as does \mathbb{Z}_n for $n = 2, 3, 4, \dots$. Because of the stipulation that $0 \neq 1$ a ring must have at least two elements and hence \mathbb{Z}_1 and \mathbb{Z}_0 are not rings.

The reader is already familiar with a variety of rings including Galois fields, as well as the fields \mathbb{Q} , \mathbb{R} , and \mathbb{C} . If F is any field, such as those just listed, then $F[X]$ is a polynomial ring over F , as is $F[X, Y]$. The Gaussian integers constitute a ring, as does each set of the form

$$\mathbb{Z}[\sqrt{d}] = \{a + b\sqrt{d} \mid a, b \in \mathbb{Z}\}$$

where d is a negative square-free integer. These rings are called *quadratic domains*. Some were discussed in the previous chapter, and others will be examined closely in this one. Their elements are also called *integers* and in order to distinguish between them and the classical integers $0, \pm 1, \pm 2, \pm 3, \dots$ these latter will be referred to as *rational integers* as opposed to the *complex integers*, or *irrational integers* that constitute the quadratic domains.

The element a of the ring R is said to be a unit of R if there exists an element $b \in R$ such that $ab = 1$. The sets of units of the rings we encountered before have been small commutative groups, but that is no longer the case. While the units always form a commutative group, their number may be infinite.

We now go on to derive some of the elementary properties of rings.

Proposition 14.1 If a is an element of the ring R , then $a \cdot 0 = 0$.

Proof. First,

$$a \cdot 0 = a \cdot (0 + 0) = a \cdot 0 + a \cdot 0.$$

If we now add $-(a \cdot 0)$ to both sides of the equation above, then we have $0 = a \cdot 0$. ■

Proposition 14.2 For any elements a and b of the ring R ,

- (a) $(-1) \cdot a = -a$;
- (b) $(-a) \cdot b = a \cdot (-b) = -(a \cdot b)$;
- (c) every element of R has at most one additive inverse;
- (d) every element of R has at most one multiplicative inverse.

Proof. See Exercises 14.1.4 to 14.1.7. ■

In any product $a \cdot b$ it is customary to omit the dot and simply write ab . An element a of a ring is said to be *irreducible* if it is a nonunit and in any factorization $a = bc$, either

b or c must be a unit. This seems to be a reasonable enough extension of the notion of primality from the integer ring \mathbb{Z} to arbitrary rings, but, for reasons that probably have to do with 20/20 hindsight, Kummer felt that it was another definition that caught the true essence of primality (see Section 12.6). This stands in marked contrast to the definition used by all mathematicians from Euclid circa 300 BCE to Gauss only 10 years earlier, that a (positive) prime number is one that is only divisible by 1 and itself. An element p of a ring is *prime* if whenever $p \mid ab$, then either $p \mid a$ or $p \mid b$. It follows from Lemma 4.6 that the primes of the ring of integers \mathbb{Z} are also prime in this sense. Exercise 6.3.15 demonstrates this for the polynomial ring $F[x]$ over any field F . Generally speaking, the notions of irreducibility and primeness are, to some extent, independent of each other. The element 2 in \mathbb{Z}_6 is prime because every even entry in the multiplication table of Table 4.3 must belong to either an even-numbered column or an even-numbered row. On the other hand, 2 is not irreducible because $2 \equiv 2 \times 4 \pmod{6}$. The converse situation cannot happen (see Exercise 14.1.1).

Exercises 14.1

1. Prove that the rational integer x is prime if and only if it is irreducible.
2. Let $n > 1$ and $k \geq 1$ be positive rational integers with the property $P(n, k)$: For every $a, b \in \mathbb{Z}$, if $n \mid a^k b^k$, then either $n \mid a^k$ or $n \mid b^k$.
 - (a) Prove that if $P(n, 1)$ is true, then there exists a rational prime p such that $n \in \{1, p\}$.
 - (b) Prove that if $P(n, 2)$ is true, then there exists a rational prime p such that $n \in \{1, p, p^2\}$.
 - (c) Prove that if $P(n, 1)$ is true, then there exists a rational prime p such that $n \in \{1, p, p^2, \dots\}$.
3. The positive rational integer n is divisible by a square if and only if there exists a rational integer m such that $n \nmid m$ and $n \mid m^2$.
4. In the proof of Proposition 14.2, prove that $(-1) \cdot a = -a$.
5. In the proof of Proposition 14.2, prove that $(-a) \cdot b = a \cdot (-b) = -(a \cdot b)$.
6. In the proof of Proposition 14.2, prove that every element of R has at most one additive inverse.
7. In the proof of Proposition 14.2, prove that every element of R has at most one multiplicative inverse.

For each of the Exercises 14.1.8 to 14.1.20 specify a set R and two binary operations on R . Decide which of these determine a ring and which do not. Identify the units. Justify your answers.

8. $\mathbb{Z}_2, +, \cdot$
9. $\mathbb{Z}_3, +, \cdot$
10. $\mathbb{Z}_4, +, \cdot$
11. $\mathbb{Z}_6, +, \cdot$
12. $\mathbb{R}, +, \cdot$
13. \mathbb{R} , subtraction, multiplication
14. \mathbb{R}^3 , vector addition, dot product
15. \mathbb{R}^3 , vector addition, cross product
16. All the subsets of \mathbb{Z} , union, intersection
17. All the subsets of \mathbb{Z} , union, Δ , where $\Delta(A, B) = (A \cap B^c) \cup (B \cap A^c)$
18. $\mathbb{Z}, -, \cdot$
19. $\{f : \mathbb{R} \rightarrow \mathbb{R}\}, +, \cdot$
20. $\{f : \mathbb{R} \rightarrow \mathbb{R} \mid f \text{ is continuous everywhere}\}, +, \cdot$

14.2 Ideals

An *ideal* of the ring R is a subset I of R such that for any $a, b \in I$ and $r \in R$

$$a \pm b \in I \quad \text{and} \quad ra \in I.$$

Two ideals are said to be *equal* if they are equal as sets. The prototypical example for an ideal is the set of even numbers in the rational integers \mathbb{Z} . More generally, if a is any element of the ring \mathbb{Z} , then

$$\langle a \rangle = \{ ra \mid r \in \mathbb{Z} \} = \{ 0, \pm a, \pm 2a, \pm 3a, \dots \}.$$

More generally yet, if the ring \mathbb{Z} is replaced by an arbitrary ring R , then

$$\langle a \rangle = \{ ra \mid r \in R \}$$

is an ideal of R . Ideals of the type displayed above are said to be *generated* by a and are called *principal ideals*. In particular, $\langle 1 \rangle = R$ which is referred to as the *unit ideal*. This observation generalizes to the next proposition.

Proposition 14.3 If a is an element of the ring R , then $\langle a \rangle = R$ if and only if a is a unit of R .

Proof. Suppose first that $\langle a \rangle = R$. It follows that there is an element $b \in R$ such that $ab = 1$. This means that a is invertible and hence it is a unit.

Conversely, if a is a unit of R , it has an inverse a^{-1} . So if r is any element of R , then $r = (aa^{-1})r = a(a^{-1}r) \in \langle a \rangle$. Thus, $\langle a \rangle = R$. ■

Let $S = \{a_1, a_2, \dots, a_n\}$ be a finite subset of the ring R . The *ideal generated by S* is the following subset of R :

$$\left\{ r = r_1 a_1 + r_2 a_2 + \dots + r_n a_n \mid r_1, r_2, \dots, r_n \in R \right\}$$

where n is any positive integer. This is denoted as

$$\langle a_1, a_2, \dots, a_n \rangle.$$

For example, $\{1, \sqrt{-1}\}$ generates $\mathbb{Z}[i]$.

In \mathbb{Z} , $\langle 2, 3 \rangle = \langle 1 \rangle$, $\langle 4, 6 \rangle = \langle 2 \rangle$ and, in general, $\langle a, b \rangle = \langle g \rangle$ where g is the greatest common divisor (a, b) . (See Exercise 14.2.6.)

An ideal A is said to be *divisible* by an ideal B if there exists an ideal C such that $A = BC$. The unit ideal $\langle 1 \rangle$ clearly consists of the whole ring R and it follows from the equation $I \cdot \langle 1 \rangle = I$ that every ideal in R is divisible by $\langle 1 \rangle$. It will prove convenient to know that the unit ideal of a ring is characterized by this property:

Proposition 14.4 If the ideal A is divisible by the ideal C , then $A \subseteq C$.

Proof. Suppose $A = CD$ where $C = \langle c_1, c_2 \rangle$ and $D = \langle d_1, d_2 \rangle$ so that

$$A = CD = \langle c_1 d_1, c_2 d_1, c_1 d_2, c_2 d_2 \rangle.$$

By the defining properties of ideals each of the four products above belongs to C and so $A \subseteq C$. ■

An ideal that is different from R and is only divisible by the unit ideal R and itself is said to be an *irreducible ideal*. An ideal that is not *irreducible* is *composite*.

We now show that the apparent paradox of the equation

$$2 \cdot 3 = (1 + \sqrt{-5})(1 - \sqrt{-5}) \quad (14.5)$$

can be resolved by replacing each of the four integers of Equation 14.5 with the principal ideal it generates and viewing them as elements of the set of all ideals of $\mathbb{Z}[\sqrt{-5}]$. In that context (see Exercise 14.2.6)

$$\langle 2 \rangle = \langle 2, 1 + \sqrt{-5} \rangle^2, \quad (14.6)$$

$$\langle 3 \rangle = \langle 3, 1 + \sqrt{-5} \rangle \langle 3, 1 - \sqrt{-5} \rangle, \quad (14.7)$$

$$\langle 1 + \sqrt{-5} \rangle = \langle 2, 1 + \sqrt{-5} \rangle \langle 3, 1 + \sqrt{-5} \rangle, \quad (14.8)$$

$$\langle 1 - \sqrt{-5} \rangle = \langle 2, 1 + \sqrt{-5} \rangle \langle 3, 1 - \sqrt{-5} \rangle. \quad (14.9)$$

Note that the factors of the right-hand sides of the above equations are, by Exercise 14.2.6, all prime ideals. When the four integers in Equations 14.6 to 14.9 are replaced by these respective right-hand sides we obtain the following:

$$\langle 6 \rangle = \langle 2 \rangle \langle 3 \rangle = \langle 2, 1 + \sqrt{-5} \rangle^2 \langle 3, 1 + \sqrt{-5} \rangle \langle 3, 1 - \sqrt{-5} \rangle,$$

and

$$\begin{aligned} \langle 6 \rangle &= \langle 1 + \sqrt{-5} \rangle \langle 1 - \sqrt{-5} \rangle \\ &= \langle 2, 1 + \sqrt{-5} \rangle \langle 3, 1 + \sqrt{-5} \rangle \langle 2, 1 - \sqrt{-5} \rangle \langle 3, 1 - \sqrt{-5} \rangle \\ &= \langle 2, 1 + \sqrt{-5} \rangle^2 \langle 3, 1 + \sqrt{-5} \rangle \langle 3, 1 - \sqrt{-5} \rangle. \end{aligned}$$

In other words, the distinct factorizations of the integer 6 above lead to one and the same factorization of the ideal $\langle 6 \rangle$ into prime ideals.

Exercises 14.2

1. Show that

$$(a) \quad \langle 3, 1 + 2\sqrt{-5} \rangle \langle 3, 1 - 2\sqrt{-5} \rangle = \langle 3 \rangle$$

$$(b) \quad \langle 7, 1 - 2\sqrt{-5} \rangle \langle 7, 1 + 2\sqrt{-5} \rangle = \langle 7 \rangle$$

$$(c) \quad \langle 3, 1 + 2\sqrt{-5} \rangle \langle 7, 1 + 2\sqrt{-5} \rangle = \langle 1 + 2\sqrt{-5} \rangle$$

$$(d) \quad \langle 3, 1 - 2\sqrt{-5} \rangle \langle 7, 1 - 2\sqrt{-5} \rangle = \langle 1 - 2\sqrt{-5} \rangle$$

2. Prove that the product of two ideals is well defined. In other words, show that if $A = A'$ and $B = B'$, then $AB = A'B'$.

3. Find an ideal X of \mathbb{Z} such that
 - (a) $\langle 6 \rangle X = \langle 18 \rangle$
 - (b) $\langle 18 \rangle X = \langle 6 \rangle$
4. Find an ideal X of $\mathbb{Z}[\sqrt{-1}]$ such that
 - (a) $\langle 2 + i \rangle X = \langle 7 - 6i \rangle$
 - (b) $\langle 7 - 6i \rangle X = 2 + i$
5. Let R be a quadratic domain and I an ideal that divides every ideal of R . Prove that $I = R$.
6. Prove Equations 14.6 to 14.9.

14.3 Domains

Now that we know that some rings do not have the property of unique factorization, it should be of interest to devise necessary and sufficient conditions that will easily (or maybe not so easily) determine whether a given ring has unique factorization. Unfortunately, this is not possible yet and we have to make do with conditions that are necessary or sufficient. In the last section we demonstrated that certain rings ($\mathbb{Z}[\sqrt{-3}]$, $\mathbb{Z}[\sqrt{-5}]$, $\mathbb{Z}[\sqrt{-7}]$) allowed for distinct factorizations. Now we give some conditions that guarantee uniqueness. Some of the results of this section will make use of ideals.

Consider the multiplication table of \mathbb{Z}_6 (Table 4.3). Observe that for every positive integer n ,

$$3^n \equiv 3 \pmod{6}.$$

The element 3 is not invertible and hence it cannot be considered to be a unit. On the other hand, if 3 were a prime, then this would show that the numbers of \mathbb{Z}_6 do not have unique factorization. Nor does 3 behave like a composite number, since it cannot be expressed as the product of two smaller numbers. The root of the problem is that

$$2 \cdot 3 \equiv 0 \pmod{6},$$

or, in other words, the product of two nonzero numbers can be zero. Such a troublesome nonzero number whose multiples do contain a 0 is called a *zero divisor*. A ring that is free of zero divisors is an *integral domain*.

Most of the rings we have encountered were integral domains. These include

- the integers \mathbb{Z} ;
- all fields, including \mathbb{Q} , \mathbb{R} , \mathbb{C} , and \mathbb{Z}_p , where p is a rational prime;
- the polynomial ring $F(X)$ where F is an arbitrary field; and
- $\mathbb{Z}[d]$.

On the other hand, if n is any composite rational integer, then \mathbb{Z}_n has zero divisors, as does the ring of continuous real-valued functions $F : [0, 1] \rightarrow \mathbb{R}$.

Let D be an integral domain. A *Euclidean function* $E : D \rightarrow \mathbb{Z}$ has the following properties:

- (a) If $p \neq 0$ and $q \neq 0$ are ring elements, then $0 \leq E(p) \leq E(pq)$.
- (b) If $p \neq 0$ and $q \neq 0$ are ring elements such that $E(q) \leq E(p)$, then there exist two other elements d and r such that $p = dq + r$ and either $r = 0$ or $E(r) < E(q)$.

An integral domain is said to be a *Euclidean domain* provided it has a Euclidean function.

It is easy to see that \mathbb{Z} is a Euclidean domain. We need merely define $E(a) = a^2$ and recognize that given the p and the q of requirement (b) above, we let d be the quotient and r be the remainder when p is divided by q . Similarly, the ring of polynomials $\mathbb{R}[x]$ is also a Euclidean domain. This time we set $E(p) = \deg(p)$. Once again properties (a) and (b) are easily seen to be satisfied where q and r are, respectively, the quotient and the remainder when p is (long) divided by q .

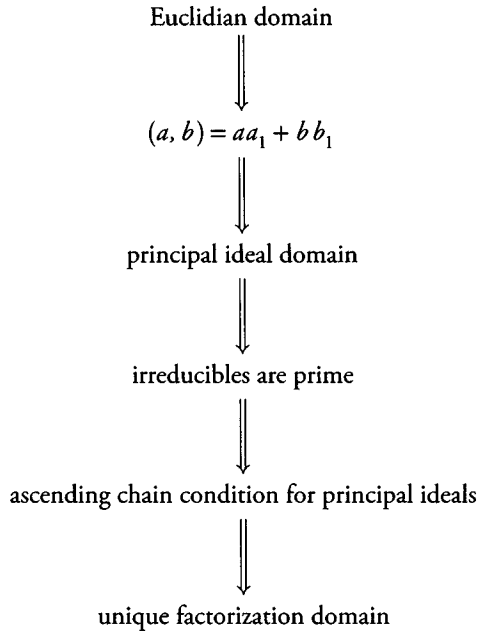
Contrariwise, $3 \in \mathbb{Z}[\sqrt{-5}]$ is not prime because

$$(2 + \sqrt{-5})(2 - \sqrt{-5}) = 3 \cdot 3$$

and yet 3 divides neither $2 + \sqrt{-5}$ nor $2 - \sqrt{-5}$ (why not?). On the other hand, a norm argument demonstrates that 3 is irreducible.

Notwithstanding these examples there is a relationship between these two notions that will be explored below.

We now set out to prove that every Euclidean domain has the unique factorization property. The proof is broken down into five parts whose relationships are depicted in the chart below. Each item of this list will be shown to be justified by the previous ones, resulting in the conclusion that every Euclidean domain has the unique factorization property.



Proposition 14.10 Let a and b be two elements of a Euclidean domain R . Then a and b have a greatest common divisor (a, b) which is expressible as a combination of a and b . That is, there exist two ring elements a_1 and b_1 such that $(a, b) = a_1 a + b_1 b$.

Proof. The proof is essentially the same as that used in the proofs of Propositions 4.1 and 6.15 and the details are omitted. One need merely create a chain like those in Figure 6.1 and note that the eventual termination of a such chain is guaranteed by the fact that the non-negative integer-valued norm of the remainder gets smaller with each application of the Euclidean algorithm. ■

A ring is a *principal ideal domain* (PID) if it has no zero divisors and all of its ideals are principal.

Proposition 14.11 Every Euclidean domain is a principal ideal domain.

Sketch of proof. Let I be an ideal of the ring R , let E be the Euclidean function of R , and let $M = \min\{E(a) \mid a \in I, a \neq 0\}$. Let $q \in I$ be an element such that $E(q) = M$. We will show, by contradiction, that $I = \langle q \rangle$. Since $q \in I$ it follows that $I \supseteq \langle q \rangle$. The reverse inclusion is proved by contradiction. Suppose $I \neq \langle q \rangle$; then there exists an element $p \in I$ such that $p \notin \langle q \rangle$. By the definition of Euclidean domain, there exist $d, r \in R$ such that $p = dq + r$ or $r = p - dq$ where $E(r) < E(q) = M$. Since q is in I so is

dq . Since $p \in I$, so is $r = p - qd$. Thus r is an element of I such that $E(r) < M$. Of necessity $r = 0$ so that $p = dq$ which of course means that $p \in \langle q \rangle$, leading to the desired contradiction. Hence $I = \langle q \rangle$. ■

Proposition 14.12 In a Euclidean (principal ideal) domain irreducible elements are primes.

Proof. Let a be an irreducible element of the PID D and suppose $a \mid bc$. We must show that either $a \mid b$ or $a \mid c$.

Consider the ideal $I = \{ax + by \mid x, y \in D\}$. By the previous proposition I is principal and so we may assume that $I = \langle d \rangle$. Since $a \in I$, we can write $a = dr$ for some r in D . Because a is irreducible either d or r is a unit.

If d is a unit, then $I = D$, and hence there exist x and y such that $1 = ax + by$, or $c = acx + bcy$. Since a divides both summands of the right-hand side, it follows that $a \mid c$.

On the other hand, if r is a unit, then $\langle a \rangle = \langle d \rangle = I$. Because $b \in I$, there is an element $t \in D$ such that $at = b$. Hence, $a \mid b$. ■

A sequence I_k of ideals is said to be an *ascending chain of ideals* if it satisfies the inclusions

$$I_k \subset I_{k+1}, \quad \text{for all } k = 1, 2, 3, \dots$$

For example, the sequence I_n , $n = 1, 2, 3, \dots$, where

$$I_n = \left\langle \frac{1}{2^n} \right\rangle$$

is an ascending chain. So is the sequence of polynomial rings $\mathbb{Z}[x_1, x_2, \dots, x_n]$.

Lemma 14.13 If $\{I_n\}$, $n = 1, 2, 3, \dots$, is an ascending sequence of ideals of some domain D , then

$$I_\infty = \bigcup_{n=1}^{\infty} I_n$$

is also an ideal of D .

Proof. Let $\alpha \in I_\infty$. Then there exists an index, call it n_α , such that $\alpha \in I_{n_\alpha}$. Note that $\alpha \in I_n$ for all $n \geq n_\alpha$ because the given sequence is increasing. For any $\beta \in I_\infty$, $\alpha, \beta \in I_{n_{\alpha+\beta}}$, and hence

$$\alpha + \beta, \alpha\beta \in I_{n_{\alpha+\beta}} \subset I_\infty.$$

Thus, the set I_∞ is closed with respect to addition and multiplication. Similar arguments can be used to prove that this set is indeed an ideal of D . ■

Proposition 14.14 Let $I_1 \subseteq I_2 \subseteq I_3 \subseteq \cdots \subseteq I_n \subseteq \cdots$ be an increasing sequence of principal ideals of the Euclidean domain D . Then there exists an integer k such that $I_n = I_k$ for all $n \geq k$.

Proof. By Lemma 14.13 it is known that

$$I_\infty = \bigcup_{n=1}^{\infty} I_n$$

is also an ideal. By Proposition 14.11, I_∞ is also principal and hence $I_\infty = \langle q \rangle$ for some $q \in D$. By the definition of the union operation, $q \in I_k$ for some $k = k_q$. Because the given sequence of ideals is increasing, it follows that

$$q \in I_n \quad \text{for all } n \geq k_q.$$

Thus, for all $n \geq k$

$$I_\infty = \langle q \rangle = I_{k_q} \subseteq I_\infty$$

and hence $n \geq k_q$ implies that $I_\infty = I_{k_q}$. ■

A *prime factorization* of an element a of a ring is an equation

$$a = \varepsilon p_1^{r_1} p_2^{r_2} \cdots p_h^{r_h} \tag{14.15}$$

where ε is a unit of R , and for each $i = 1, 2, 3, \dots, h$, the integer p_i is a prime element of R , and r_i is a nonnegative rational integer. A ring is said to have the *unique factorization property* if given two prime factorizations of associate elements such as Equation 14.15 and

$$a' = \delta q_1^{s_1} q_2^{s_2} \cdots q_k^{s_k} \tag{14.16}$$

necessarily $h = k$ and there is a relabeling of the q_i 's such that p_i and q_i are associates for each $i = 1, 2, 3, \dots, h = k$.

Proposition 14.17 Every Euclidean domain has the unique factorization property.

Proof. We first show, by contradiction, that every element $a \in D$ has a factorization into irreducible elements. Suppose that is not the case for the element a_0 . Then a_0 cannot be irreducible for if it were, then a_0 would be a factorization into irreducibles for itself.

Since a_0 is decomposable, there exist nonunits a_1 and b_1 such that $a_0 = a_1 b_1$. By assumption, either a_1 or b_1 is decomposable; relabel if necessary so that a_1 is reducible. Hence there exist nonunits a_2 and b_2 such that $a_1 = a_2 b_2$. Once again, the hypothesis implies that at least one of a_2 and b_2 is reducible. Relabel, if necessary, so that a_2 is reducible. Continuing indefinitely we produce a sequence of distinct nonunits a_0, a_1, a_2, \dots such that $a_{n+1} \mid a_n$. It follows that we have an increasing sequence of ideals

$$\langle a_0 \rangle \subsetneq \langle a_1 \rangle \subsetneq \langle a_2 \rangle \subsetneq \langle a_3 \rangle \subsetneq \dots$$

which clearly does not stabilize. This contradicts the previous proposition and so there must exist a factorization of a_0 into irreducible elements. By Proposition 14.12 this is also a prime factorization of a_0 . To prove that Euclidean domains have unique factorization, the induction process of Theorem 4.9 must be modified. For any factorization

$$\varphi: a = \varepsilon p_1^{r_1} p_2^{r_2} \cdots p_h^{r_h}$$

we define the length of the factorization φ to be $\lambda(\varphi) = r_1 + r_2 + \cdots + r_h$. We proceed to prove the uniqueness by induction on λ . Suppose the element a has a factorization of length zero. Then a is a unit and every factorization of a has length zero and the unique factorization holds. Let n be a positive rational integer and suppose that the unique factorization holds for all the elements of R that have a factorization of length less than n . Let a be a ring element with a factorization

$$\varphi: a = \delta p_1^{r_1} p_2^{r_2} \cdots p_h^{r_h}$$

of length n . Let

$$\psi: a = \varepsilon q_1^{s_1} q_2^{s_2} \cdots q_k^{s_k}$$

be another factorization of a . Since $p_1 \mid a$ it follows from the definition of primality that $p_1 \mid q_i$ for some $i \in \{1, 2, \dots, k\}$. Because D is a Euclidean domain both p_1 and q_i are irreducibles and hence there exists a unit η so that $p_1 = \eta q_i$, or, after relabeling, $p_1 = \eta q_1$. Let φ' and ψ' be the factorizations obtained when φ and ψ are divided by p_1 and ηq_1 , respectively. Then

$$\varphi': \frac{a}{p_1} = \delta p_1^{r_1-1} p_2^{r_2} \cdots p_h^{r_h}$$

and

$$\psi' : \frac{a}{\eta q_1} = \varepsilon q_1^{s_1-1} q_2^{s_2} \cdots q_k^{r_k}.$$

Since $\lambda(\psi') = \lambda(\psi) - 1 < n$, the induction hypothesis (that a has unique factorization) applies to a/p_1 so that $h = k$, p_i and q_i are associates, and $r_i = s_i$ for all $i = 1, 2, \dots, h = k$. Hence φ and ψ are the same factorizations. This completes the induction and the proof. ■

Corollary 14.18 The Gaussian and Eulerian numbers have the unique factorization property.

Exercises 14.3

1. Prove that every finite integral domain is a field.

14.4 Quotients of Rings

By definition, every ring R is a commutative group with respect to its “+” operation, and every ideal I of R constitutes a subgroup of R with respect to this addition operation. As we saw in Chapter 10, this set of circumstances yields a quotient structure R/I whose elements are cosets $a + I$. Since the addition on R is commutative, we needn’t be concerned about right cosets versus left cosets—they are the same. As noted in Theorem 10.2, the addition of cosets, defined by

$$(a + I) + (b + I) = (a + b) + I,$$

defines a group R/I whose elements are the cosets of I and whose binary operation is defined above. With minor changes the statement of the Law of Homomorphisms holds in the context of rings as well.

Theorem 14.19 (The Law of Homomorphisms) Let $f : G \rightarrow Q$ be a surjective ring homomorphism. Then $\text{Ker } f$ is a subring of G such that $G/\text{Ker } f \cong Q$.

For example, if $R = \mathbb{Z}_6$ and $I = \{0, 3\}$, then the quotient ring has as its elements the cosets $\{0, 3\}$, $\{1, 4\}$, and $\{2, 5\}$. This observation may be restated as

$$\mathbb{Z}_6 / \{0, 3\} \cong \mathbb{Z}_3.$$

For example, let $R = \mathbb{Z}[i]$ and $I = \langle 1 + i \rangle$. We will demonstrate that the quotient group R/I is in fact isomorphic to $\text{GF}(2, x)$.

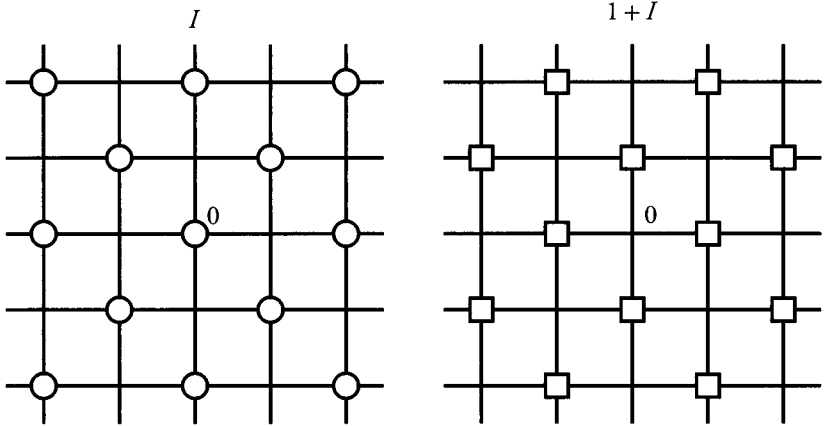


Figure 14.1 Two interlaced lattices

+				
	I	$1 + I$		
I	I	$1 + I$		
$1 + I$	$1 + I$	I		

·				
	I	$1 + I$		
I	I	I		
$1 + I$	I	$1 + I$		

Figure 14.2 A ring with two elements

Note that the following statements are logically equivalent: $x + yi \in \langle 1 + i \rangle$ if and only if $x + yi = (a + bi)(1 + i)$ has an integer solution for a and b if and only if $a = (x + y)/2$ and $b = (y - x)/2$ are integers with the same parities if and only if $x \equiv y \pmod{2}$ (see Figure 14.1).

It follows that the coset

$$1 + I = \{x + iy \mid x \not\equiv y \pmod{2}\}$$

is the only other coset besides I . The addition and multiplication tables of the quotient group appear in Figure 14.2 and it is easily verified that this ring is isomorphic to the Galois field $\text{GF}(2, x)$.

Another example appears in Tables 10.4 to 10.7 wherein the Galois field $\text{GF}(2, x^2 + x + 1)$ is displayed as the quotient ring $\mathbb{Z}_2[x]/(x^2 + x + 1)$.

Given a ring R and an ideal $I \subseteq R$, the ideal is said to be a *maximal ideal* if the only ideals that contain I are itself and R .

Theorem 14.20 An ideal I of a ring R is maximal if and only if R/I is a field (with respect to the quotient operations).

Proof. Suppose I is a maximal ideal of R and let a be any ring element not in I so that $a + I \neq I$. It follows from the maximality of I that the ideal (a, I) consists of all the elements of R . In particular, there exist elements $b \in R$ and $c \in I$ such that $ab + c = 1$. Hence,

$$(a + I) \odot (b + I) = ab + I = ab + c + I = 1 + I = 1_{R/I}.$$

Conversely, suppose R/I is a field and let J be any ideal of R that properly contains I . We now show that necessarily $J = R$.

Since $J \supsetneq I$ there exists an element a such that $a \in J - I$. Consequently $a + I \neq I = 0_{R/I}$. Since R/I is a field the quotient element $a + I$ has a multiplicative inverse, say $b + I$. Thus

$$ab + I = (a + I) \odot (b + I) = 1 + I.$$

It follows that $1 - ab \in I \subseteq J$. Since $a \in J$ and J is an ideal, it follows that $ab \in J$. Thus,

$$1 = (1 - ab) + ab \in J + J = J.$$

Thus $J = R$. ■

Lemma 14.21 Let F be a field and $p(x)$ an irreducible polynomial in $F[x]$. Then $\langle p(x) \rangle$ is a maximal ideal of $F(x)$.

Proof. Let I be an ideal that properly contains $\langle p(x) \rangle$ and let $q(x)$ be another polynomial in $I - \langle p(x) \rangle$. Then, because $F[x]$ is a PID, there is a polynomial $q(x) \in R$ such that $I = \langle q(x) \rangle$. Since $p(x) \in I$ it follows that $q(x) \mid p(x)$. The irreducibility of $p(x)$ now implies that $q(x)$ is a unit and hence $I = R$. ■

Theorem 14.22 Let $p \in \mathbb{Z}$ be a prime, F be the field \mathbb{Z}_p , and $P(x)$ be a primitive polynomial in $F[x]$. Then $F[x]/\langle P(x) \rangle \simeq \text{GF}(p, P(x))$.

Proof. Let α be a primitive element of $\text{GF}(p, P(x))$ and define a function $f : F[x] \rightarrow \text{GF}(p, P(x))$ via $f(q(x)) = q(\alpha)$ where $q(x)$ is an arbitrary polynomial in $F[x]$. That f is a homomorphism of $F[x]$ follows from

$$f(q(\alpha))f(q'(\alpha)) = q(\alpha)q'(\alpha) = f(q(\alpha)q'(\alpha)).$$

Galois' Primitive Element 7.17 guarantees the surjectivity of f . Thus, f is a homomorphism of $F[x]$ whose kernel is $\langle P(x) \rangle$. The desired result follows from the Law of Homomorphisms. ■

We now turn to examine some quotient rings of the Gaussian integers $\mathbb{Z}[\sqrt{-1}]$. Given a group G and a normal subgroup H , a *complete residue system* is a set S such that every coset of H contains exactly one element of S . For example, the set $\{1, a, b, c\}$ is a complete residue system of the quotient \mathbb{Z}_{12}/H portrayed in Table 10.2 and Figure 10.2, and the set $\{d, e, b, g\}$ is a complete residue system of the subgroup H of the Quaternions (see Table 10.2 and Figure 10.2). The set $\{1, 2, 3, \dots, n\}$ is such a system for the ideal $\langle n \rangle$ in \mathbb{Z} .

For example, we will find a complete residue system for $\mathbb{Z}[i]/\langle 2+i \rangle$. Since $N(2+i) = 5$, it follows from the long division process that every coset of $\langle 2+i \rangle$ has an element of norm less than 5. Hence the set of elements in Figure 14.3, call it V , contains a complete residue system. The set V is winnowed down to a complete residue system of Figure 14.3 by the observation that if for some unit $\varepsilon \in \{\pm 1, \pm i\}$, $z_2 = z_1 + (2+i)\varepsilon$, then z_1 and z_2 belong to the same coset and so one of them can be deleted. The complete residue system clearly contains five elements and their interaction is identical with the two tables in Figure 14.4.

Given two rings R and S and a function $f: R \rightarrow S$ such that

- (a) f is injective (one-to-one),
- (b) f is onto (surjective),
- (c) $f(x+y) = f(x) + f(y)$ for all $x, y \in R$,
- (d) $f(xy) = f(x)f(y)$ for all $x, y \in R$,

the rings R and S are said to be *isomorphic*, and $f: R \rightarrow S$ is a *ring isomorphism* of R and S . This is, of course, quite similar to the definition of a group isomorphism (Section 9.3) and even more so to the definition of an isomorphism of fields (Section 10.3). In fact, it could be said that a ring isomorphism is a group isomorphism that also satisfies requirement (d) above.

A *prime subfield* of the field F is one which does not have any proper subfields.

Proposition 14.23 Every field has a unique prime subfield.

Proof. Let F be a field and let \tilde{F} denote the intersection of all the subfields of F . By Exercise 14.4.7, \tilde{F} is a field. Moreover, if $\tilde{F} \subsetneq F$ contained a proper subfield, say G , then $\tilde{F} \cap G = G \subsetneq \tilde{F}$, contradicting the definition of \tilde{F} . Hence \tilde{F} is a prime subfield of F . If H is any other prime subfield of F , then the field $H \cap \tilde{F}$ is a subfield that must equal both and hence $H = \tilde{F}$. ■

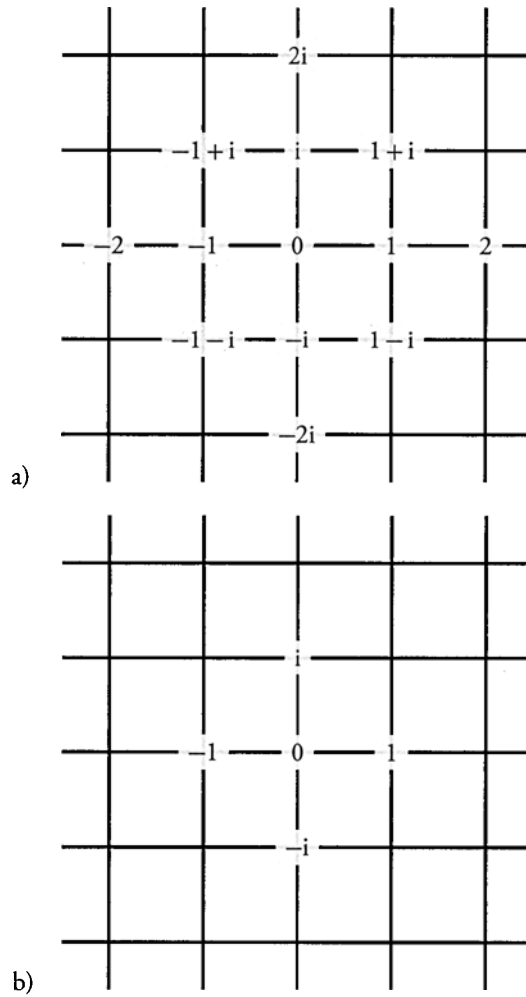


Figure 14.3 In search of a ring

+	0	1	-1	i	-i
0	0	1	-1	i	-i
1	1	-i	0	-1	i
-1	-1	0	i	-i	1
i	i	-1	-i	1	0
-i	-i	i	1	0	-1

·	0	1	-1	i	-i
0	0	0	0	0	0
1	0	1	-1	i	-i
-1	0	-1	1	-i	i
i	0	i	-i	-1	1
-i	0	-i	i	1	-1

 Figure 14.4 $\mathbb{Z}[i]/\langle 2+i \rangle \cong \mathbb{Z}_5$

Proposition 14.24 If F is an arbitrary field, then F contains a subfield that is isomorphic to exactly one of the fields

$$\mathbb{Q}, \mathbb{Z}_2, \mathbb{Z}_3, \mathbb{Z}_5, \dots, \mathbb{Z}_p, \dots$$

where p varies over all the rational primes.

Proof. Let 1_F be the multiplicative identity of F and for each positive integer n set $n \cdot 1_F$ equal to the sum of n 1_F 's. Consider the sequence

$$\Gamma = \{ n \cdot 1_F \mid n = 1, 2, 3, \dots \} \subset F.$$

The terms of this sequence are either all distinct or else there exist coefficients g and h such that $g \cdot 1_F = h \cdot 1_F$. In the first case let $f: \mathbb{Q} \rightarrow F$ be defined as follows. First set $f(m) = m \cdot 1_F$ for $m \in \mathbb{Z}$ and

$$f(m/n) = (m \cdot 1_F)(n \cdot 1_F)^{-1} \in F.$$

The proofs of the following facts are straightforward (see Exercise 14.4.8):

- (a) The function f is well defined.
- (b) The range of f is in F .
- (c) The function f is injective.
- (d) $f(x + y) = f(x) + f(y)$ and $f(xy) = f(x)f(y)$.

It follows that f is an isomorphism of \mathbb{Q} and there exists a subset $B \subset F$ such that $\Gamma \subset B \subset F$. In the second case observe that $(g - h) \cdot 1_F = 0$. Set

$$G = \{ g \in \mathbb{Z} \mid g \cdot 1_F = 0 \}.$$

Since G is an ideal of \mathbb{Z} , it must be principal and hence there exists a positive integer p such that $G = \langle p \rangle$. Once again we define $f: \mathbb{Z}_p \rightarrow F$ by $f(m) = m \cdot 1_F$ for $m \in \mathbb{Z}_p$, and leave it for the reader to verify that the above four properties again hold for F . Hence f is an isomorphism of \mathbb{Z}_p with a subset of F .

The reason p must be a prime is that \mathbb{Z}_p is isomorphic to a subset of F , which, being a field, has no zero divisors. The uniqueness of p follows from Proposition 14.23. ■

Let $f: \mathbb{Z} \rightarrow \mathbb{Z}_3$ be defined via

$$f(x) = \bar{x} \in \{0, 1, 2\}, \quad \bar{x} \equiv x \pmod{3}.$$

Let $R = \mathbb{Z}[i]$ and let $I = \langle 3 \rangle$. Define the function $f : \mathbb{Z}[i] \rightarrow \mathbb{Z}_3$ via

$$f(x) = \bar{x} \in \{0, 1, 2\}, \quad \bar{x} \equiv x \pmod{3}.$$

Then f is a surjective homomorphism from R onto \mathbb{Z}_3 . The kernel of this homomorphism is $\{x \in \mathbb{Z} \mid 3 \mid x\}$ or $\langle 3 \rangle$.

Let $f : \mathbb{Z}[i] \rightarrow \mathbb{Z}$ be defined as

$$f(x + iy) = \bar{x} \in \{0, 1, 2\}, \quad \bar{x} \equiv x \pmod{3}.$$

Then

$$\text{Ker } f = \{x + iy \mid \bar{x} \equiv 0 \pmod{3}\}.$$

Let $f(x + iy) = x + y$. Then

$$\text{Ker } f = \{x + iy \in \mathbb{Z} \mid x + y = 0\} = \{x + iy \in \mathbb{Z} \mid y = -x\} = \langle 1 - i \rangle.$$

Exercises 14.4

1. Prove that for any two positive integers m and n , $\mathbb{Z}_{mn}/\langle m \rangle \cong \mathbb{Z}_n$.
2. Let α be any Gaussian integer. Prove that $\mathbb{Z}[i]/\langle \alpha \rangle$ is finite.
3. Find a complete residue system for $\langle 3 + i \rangle$ in $\mathbb{Z}[i]$. Use the residue system to create both an addition and a multiplication table for $\mathbb{Z}[i]/\langle 3 + i \rangle$. Is this a field? Justify your answer.
4. Find a complete residue system for $\langle 3 + 2i \rangle$ in $\mathbb{Z}[i]$. Use the residue system to create both an addition and a multiplication table for $\mathbb{Z}[i]/\langle 3 + 2i \rangle$. Is this a field? Justify your answer.
5. Let F denote the set of all the continuous functions $f : [0, 1] \rightarrow \mathbb{R}$.
 - (a) Prove that the subset of all the functions $f \in F$ such that $f(0) = 0$ is a maximal ideal of F .
 - (b) Find four other maximal ideals of F .
 - (c) Find three ideals I_1, I_2, I_3 of F such that $I_1 \supsetneq I_2 \supsetneq I_3$.

6. Let R be the ring with underlying set

$$\left\{ \begin{pmatrix} a & b \\ 0 & a \end{pmatrix} \mid a, b \in \mathbb{R} \right\}$$

whose addition and multiplication are the standard matrix operations.

- (a) Explain why these operations define a ring on R .
 (b) Explain why the subset

$$I = \left\{ \begin{pmatrix} 0 & b \\ 0 & 0 \end{pmatrix} \mid b \in \mathbb{R} \right\}$$

is an ideal of R .

- (c) Prove that $R/I \cong \mathbb{R}$.
 7. Prove that the intersection of all the subfields of a fixed field F is a field.
 8. Complete the proof of Proposition 14.24.

Chapter Summary

Rings and ideals were defined abstractly. However, we met an old friend—the quotient structures that were so useful in the context of group theory turn out to provide just the right language for the rings and ideals as well.

Chapter Review Exercises

Mark the following true or false.

1. Every ideal is principal.
2. Every ring has at least two distinct ideals.
3. Every ring has a nonprincipal ideal.
4. If the ideal I of the ring R is maximal, then R/I is a domain.
5. If the ideal I of R is such that R/I is a domain, then R/I is a field.
6. The Gaussian integers have unique factorization.
7. The Eulerian integers have unique factorization.
8. The ring $\mathbb{Z}[\sqrt{-2}]$ has unique factorization.
9. There exists a field with exactly 6 elements.
10. For each positive integer n greater than 1, there exists a ring of order n .

New Terms

ascending chain of ideals, 364
 complete residue system, 370
 complex integers, 356
 Euclidean domain, 362
 Euclidean function, 362
 ideal, 358
 integers, 356
 integral domain, 361
 irrational integers, 356
 irreducible, 356
 irreducible ideal, 359
 isomorphic, 370
 maximal ideal, 368
 prime, 357
 prime factorization, 365
 prime subfield, 370
 principal ideal domain, 363
 principal ideals, 358
 quadratic domains, 356
 rational integers, 356
 ring, 355
 ring isomorphism, 370
 unique factorization property, 365
 unit ideal, 358
 zero divisor, 361

Supplementary Exercises

1. How many rings are there of order n where $n = 1, 2, 3$?
2. Is there a noncommutative ring of order 4?
3. Is there a noncommutative ring of order 5?

A. Excerpts from Al-Khwarizmi's *Solution of the Quadratic Equation*¹

Containing Demonstrations of the Rules of the Equations of Algebra.

...furthermore I discovered that the numbers of restoration and opposition are composed of these three kinds: namely, roots, squares, and numbers.² However, number alone is connected neither with roots nor with squares by any ratio. Of these, then, the root is anything composed of units which can be multiplied by itself, or any number greater than unity multiplied by itself: or that which is found to be diminished below unity when multiplied by itself. The square is that which results from the multiplication of a root by itself.

Of these three forms, then, two may be equal to each other, as for example:

Squares equal to roots,

Squares equal to numbers, and

Roots equal to numbers.³

Chapter I. Concerning Squares Equal to Roots⁴

The following is an example of squares equal to roots: a square is equal to five roots. The root of the square then is five, and 25 forms its square which, of course, equals five of its roots.

Another example: the third part of a square equals four roots. Then the root of the square is 12 and 144 designates its square. And similarly, five squares equal ten roots. Therefore, one square equals two roots and the root of the square is two. Four represents the square.

¹Cardano, G., *Ars Magna or The Rules of Algebra*, T. Richard Witmer translator. 1968. Reprinted by permission of MIT press

²The term "roots" (*radices*) stands for multiples of the unknown, our x ; the term "squares" (*substantiae*) stands for multiples of our x^2 ; "numbers" (*numeri*) are constants.

³In our notation, $x^2 = ax$, $x^2 = b$, $x = c$.

⁴Latin: *de substantiis numeros coequantibus*. The examples are $x^2 = 5x$, $x^2/3 = 4x$, $5x = 10x$.

In the same manner, then, that which involves more than one square, or is less than one, is reduced to one square. Likewise you perform the same operation upon the roots which accompany the squares.

Chapter II. Concerning Squares Equal to Numbers

Squares equal to numbers are illustrated in the following manner: a square is equal to nine. Then nine measures the square of which three represents one root.

Whether there are many or few squares, they will have to be reduced in the same manner to the form of one square. That is to say, if there are two or three or four squares, or even more, the equation formed by them with their roots is to be reduced to the form of one square with its root. Further, if there be less than one square, that is, if a third or a fourth or a fifth part of a square or root is proposed, this is treated in the same manner.

For example, five squares equal 80. Therefore, one square equals the fifth part of the number 80, which, of course, is 16. Or, to take another example, half of a square equals 18. This square therefore equals 36. In like manner all squares, however many, are reduced to one square, or what is less than one is reduced to one square. The same operation must be performed upon the numbers which accompany the squares.

Chapter III. Concerning Roots Equal to Numbers

The following is an example of roots equal to numbers: a root is equal to three. Therefore nine is the square of this root.

Another example: four roots equal 20. Therefore one root of this square is five. Still another example: half a root is equal to ten. The whole root therefore equals 20, of which, of course, 400 represents the square.

Therefore roots and squares and pure numbers are, as we have shown, distinguished from one another. Whence also from these three kinds which we have just explained, three distinct types of equations are formed involving three elements, as

A square and roots equal to numbers,

A square and numbers equal to roots, and

Roots and numbers equal to a square.

Chapter IV. Concerning Squares and Roots Equal to Numbers

The following is an example of squares and roots equal to numbers: a square and ten roots are equal to 39 units.⁵ The question therefore in this type of equation is about as

⁵This example, $x^2 + 10x = 39$, with answer $x = 3$, "runs," as Karpinski notices in his introduction to this translation, "like a thread of gold through the algebras for several centuries, appearing in the algebras of Abu

follows: what is the square which combined with ten of its roots will give a sum total of 39? The manner of solving this type of equation is to take one-half of the roots just mentioned. Now the roots in the problem before us are ten. Therefore, take five, which multiplied by itself gives 25, an amount which you add to 39, giving 64. Having taken then the square root of this which is eight, subtract from it the half of the roots, five, leaving three. The number three therefore represents one root of this square, which itself, of course, is nine. Nine therefore gives that square.

Similarly, however many squares are proposed all are to be reduced to one square. Similarly also you may reduce whatever numbers or roots accompany them in the same way in which you have reduced the squares.

The following is an example of this reduction: two squares and ten roots equal 48 units. The question therefore in this type of equation is something like this: what are the two squares which when combined are such that if ten roots of them are added, the sum total equals 48? First of all it is necessary that the two squares be reduced to one. But since one square is the half of two, it is at once evident that you should divide by two all the given terms in this problem. This gives a square and five roots equal to 24 units. The meaning of this is about as follows: what is the square which amounts to 24 when you add to it five of its roots? At the outset it is necessary, recalling the rule above given that you take one-half of the roots. This gives two and one-half which multiplied by itself gives $6\frac{1}{4}$. Add to this 24, giving $30\frac{1}{4}$. Take then of this total the square root, which is, of course, $5\frac{1}{2}$. From this subtract half of the roots, $2\frac{1}{2}$, leaving three, which expresses one root of the square, which itself is nine.

Chapter VI. Geometrical Demonstrations⁶

We have said enough, says Al-Khowarizmi, so far as numbers are concerned, about the six types of equations. Now, however, it is necessary that we should demonstrate geometrically the truth of the same problems which we have explained in numbers. Therefore our first proposition is this, that a square and ten roots equals 39 units.

The proof is that we construct (Figure A.1) a square of unknown sides, and let this square figure represent the square (second power of the unknown) which together with its

Kamil, Al-Karkhi and Omar al-Khayyami, and frequently in the works of Christian writers," and it still graces our present algebra texts. The solution of this type, $x^2 + ax = b$, is, as we can verify, based on the formula $x = \sqrt{(a/2)^2 + b} - a/2$.

⁶For these geometric demonstrations, we must go back, as said, to Euclid's *Elements* (Book VI, Propositions 28 and 29; see also Book II, Propositions 5 and 6). See also on this subject the introduction to the *Principal works of Simon Stevin*.

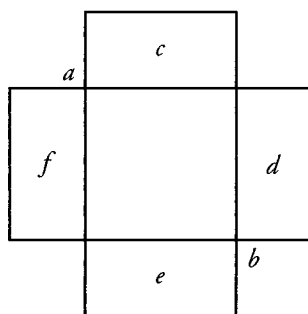


Figure A.1

root you wish to find. Let the square, then, be ab , of which any side represents one root. When we multiply any side of this by a number (or numbers) it is evident that which results from the multiplication will be a number of roots equal to the root of the same number (of the square). Since then ten roots were proposed with the square, we take a fourth part of the number ten and apply to each side of the square an area of equidistant slides, of which the length should be the same as the length of the square first described and the breadth $2\frac{1}{2}$, which is a fourth part of 10. Therefore four areas of equidistant sides are applied to the first square, ab . Of each of these the length is the length of one root of the square ab and also the breadth of each is $2\frac{1}{2}$, as we have just said. These now are the areas c , d , e , and f . Therefore it follows from what we have said that there will be four sides of unequal length, which also are regarded as unknown. The size of the areas in each of the four corners, which is found by multiplying $2\frac{1}{2}$ by $2\frac{1}{2}$, completes that which is lacking in the larger or whole area. Whence it is that we complete the drawing of the larger area by the addition of the four products, each $2\frac{1}{2}$ by $2\frac{1}{2}$; the whole of this multiplication gives 25.

And now it is evident that the first square figure, which represents the square of the unknown $[x^2]$, and the four surrounding areas $[10x]$ make 39. When we add 25 to this, that is, the four smaller squares which indeed are placed at the four angles of the square ab , the drawing of the larger square, called GH , is completed (Figure A.2). Whence also the sum total of this is 64, of which eight is the root, and by this is designated one side of the completed figure. Therefore when we subtract from eight twice the fourth part of ten, which is placed at the extremities of the larger square GH , there will remain but three. Five being subtracted from eight, three necessarily remains, which is equal to one side the first square ab .

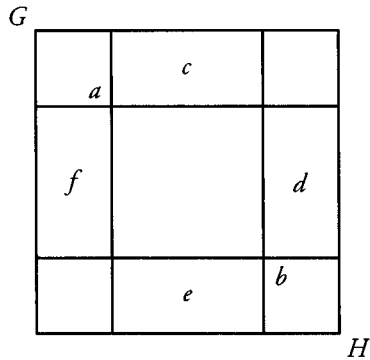


Figure A.2

This three then expresses one root of the square figure, that is, one root of the proposed square of the unknown, and nine the square itself. Hence we take half of ten and multiply this by itself. We then add the whole product of the multiplication to 39, that the drawing of the larger square GH may be completed; for the lack of the four corners rendered incomplete the drawing of the whole of this square. Now it is evident that the fourth part of any number multiplied by itself and then multiplied by four gives the same number as half of the number multiplied by itself. Therefore, if half of the root is multiplied by itself, the sum total of this multiplication will wipe out, equal, or cancel the multiplication of the fourth part by itself and then by four.

The remainder of the treatise deals with problems that can be reduced to one of the six types, for example, how to divide ten into two parts in such a way that the sum of the products obtained by multiplying each part by itself is equal to 58: $x^2 + (10 - x)^2 = 58$, so that $x = 3$ or $x = 7$. This is followed by a section on problems of inheritance.

B. Excerpts from Cardano's *Ars Magna*¹

Chapter XL On the Cube and First Power Equal to the Number

Scipio Ferro of Bologna well-nigh thirty years ago discovered this rule and handed it on to Antonio Maria Fior of Venice, whose contest with Niccolò Tartaglia of Brescia gave Niccolò occasion to discover it. He [Tartaglia] gave it to me in response to my entreaties, though withholding the demonstration. Armed with this assistance, I sought out its demonstration in [various] forms. This was very difficult. My version of it is as follows.

Demonstration

For example, let GH^3 plus six times its side GH equal 20, and let AE and CL be two cubes the difference between which is 20 and such that the product of AC , the side [of one], and CK , the side [of the other], is 2, namely one-third the coefficient of x . Marking off BC equal to CK , I say that, if this is done, the remaining line AB is equal to GH and is, therefore, the value of x , for GH has already been given as [equal to x].

In accordance with the first proposition of the sixth chapter of this book, I complete the bodies DA , DC , DE , and DF ; and as DC represents BC^3 , so DF represents AB^3 , DA represents $3(BC \times AB^2)$ and DE represents $3(AB \times BC^2)$. Since, therefore, $AC \times CD$ equals 2, $AC \times 3CK$ will equal 6, the coefficient of x ; therefore $AB \times 3(AC \times CK)$ makes $6x$ or $6AB$, wherefore three times the product of AB , AC , and BC is $6AB$. Now the difference between AC^3 and CK^3 —manifesting itself as BC^3 , which is equal to this by supposition—is 20, and from the first proposition of the sixth chapter is the sum of the bodies DA , DE , and DF . Therefore these three bodies equal 20.

Now, assume that BC is negative:

$$AB^3 = AC^3 + 3(AC \times CB^2) + (-BC^3) + 3(-BC \times AC^2),$$

¹Cardano, G., *Ars Magna or The Rules of Algebra*, T. Richard Witmer translator. 1968. Reprinted by permission of MIT press

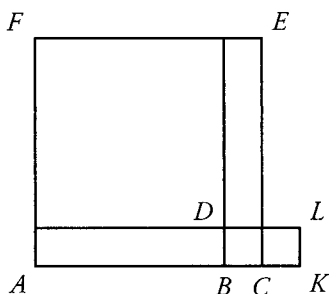


Figure B.1

by that demonstration. The difference between $3(BC \times AC^2)$ and $3(AC \times BC^2)$, however, is [three times] the product of AB , BC , and AC . Therefore, since this, as was demonstrated, is equal to $6AB$, add $6AB$ to the product of $3(AC \times BC^2)$, making $3(BC \times AC^2)$. But since BC is negative, it is now clear that $3(BC \times AC^2)$ is negative and the remainder which is equal to it is positive. Therefore,

$$3(CB \times AB^2) + 3(AC \times BC^2) + 6AB = 0.$$

It will be seen, therefore, that as much as is the difference between AC^3 and BC^3 , so much is the sum of

$$AC^3 + 3(AC \times CB^2) + 3(-CB \times AC^2) + (-BC^3) + 6AB.$$

This, therefore, is 20 and, since the difference between AC^3 and BC^3 is 20, then, by the second proposition of the sixth chapter, assuming BC to be negative,

$$AB^3 = AC^3 + 3(AC \times BC^2) + (-BC^3) + 3(-BC \times AC^2).$$

Therefore, since we now agree that

$$AB^3 + 6AB = AC^3 + 3(AC \times BC^2) + 3(-BC \times AC^2) + (-BC^3) + 6AB,$$

which equals 20, as has been proved, $[AB^3 + 6AB]$ will equal 20.

Since, therefore, $AB^3 + 6AB = 20$, and since $GH^3 + 6GH = 20$, it will be seen at once and from what is said in I-35 and XI-31 of the *Elements* that GH will equal AB . Therefore GH is the difference between AC and CB . AC and CB , or AC and CK , the coefficients, however, are lines containing a surface equal to one-third the coefficient of x and their cubes differ by the constant of the equation. Whence we have the rule:

Rule

Cube one-third the coefficient of x ; add to it the square of one-half the constant of the equation; and take the square root of the whole. You will duplicate this, and to one of the two you add one-half the number you have already squared and from the other you subtract one-half the same. You will then have a *binomium* and its *apotome*. Then, subtracting the cube root of the *apotome* from the cube root of the *binomium*, the remainder [or] that which is left is the value of x .

For example, $x^3 + 6x = 20$. Cube 2, which is one-third of 6, making 8; square 10, which is one-half the constant; 100 results. Add 100 and 8, making 108, the square root of which is $\sqrt{108}$. This you will duplicate: to one add 10, one-half the constant, and from the other subtract the same. Thus you will obtain the *binomium* $\sqrt{108} + 10$ and its *apotome* $\sqrt{108} - 10$. Take the cube roots of these. Subtract [the cube root of the] *apotome* from that of the *binomium* and you will have the value of x :

$$\sqrt[3]{\sqrt{108} + 10} - \sqrt[3]{\sqrt{108} - 10}.$$

Again, $x^3 + 3x = 10$. Cube 1, one-third of 3, and 1 results; square 5, one-half of 10, and 25 results; add 25 and 1, making 26; add 5 to and subtract it from the square root of this. You will thus form the *binomium* $\sqrt{26} + 5$ and its *apotome* $\sqrt{26} - 5$; whence x equals

$$\sqrt[3]{\sqrt{26} + 5} - \sqrt[3]{\sqrt{26} - 5}.$$

Here you have proof:

	$\sqrt[3]{\sqrt{26}+5}$ and $-\sqrt[3]{\sqrt{26}-5}$
<hr/>	
The cubes of the parts (as is evident, the sum of these is 10):	$\sqrt{26}+5$ and $-(\sqrt{26}-5)$
The squares of the parts:	$\sqrt[3]{51+\sqrt{2,600}}$ and $\sqrt[3]{51-\sqrt{2,600}}$
Three times the squares of the parts:	$\sqrt[3]{1,377+\sqrt{1,895,400}}$ and $\sqrt[3]{1,377-\sqrt{1,895,400}}$
The parts themselves:	$-\sqrt[3]{\sqrt{26}-5}$ and $\sqrt[3]{\sqrt{26}+5}$
The products of the parts and three times their squares:	$\sqrt[3]{\sqrt{49,299,354}+6,885-\sqrt{47,385,000}-7,020}$ and $-\sqrt[3]{\sqrt{49,299,354}-6,885-\sqrt{47,385,000}+7,020}$

Moreover, the cube roots contain four terms which can be reduced to two, for when 6,885 is subtracted from 7,020, the remainder is 135, and likewise when $\sqrt{47,385,000}$ is subtracted from $\sqrt{49,299,354}$ there is left $\sqrt{18,954}$. Therefore these products are

$$\sqrt[3]{\sqrt{18,954}-135}-\sqrt[3]{\sqrt{18,954}+135}.$$

The whole cube, then, from the demonstration in the third book is

$$10+\sqrt[3]{\sqrt{18,954}-135}-\sqrt[3]{\sqrt{18,954}+135},$$

and three times the root, or $3x$, equals

$$\sqrt[3]{\sqrt{18,954}-135}-\sqrt[3]{\sqrt{18,954}+135}.$$

Finally, having added all together, since the universal cube roots cancel each other, the whole becomes x^3+3x which equals exactly 10.

A third example: $x^3+6x=2$. Raise 2, one-third the coefficient of x , to the cube and 8 is the result; square 1, half of 2, making 1; add 8 to 1, and 9 is produced, the square

root of which is 3. Now duplicate 3 and to one add 1, half the constant, thus making 4, and from the other subtract half the constant, thus making 2. Then subtract the cube root of the less from the cube root of the greater and you have $\sqrt[3]{4} - \sqrt[3]{2}$ as the value of x .

Remember what we said in the chapter in the third book on extracting cube roots whenever these universal cube roots are equivalent to a whole number or a fraction. Thus in the first example,

$$\sqrt[3]{\sqrt{108} + 10} - \sqrt[3]{\sqrt{108} - 10}$$

is 2, as is indicated by the rule there given and as is perfectly clear if it is tried out.

C. Excerpts from Abel's *A Demonstration of the Impossibility of the Algebraic Resolution of General Equations Whose Degree Exceeds Four*¹

As is well known, it is possible to resolve the general equations up to the fourth degree, but the equations of higher degree only in some special cases, and, if I am not mistaken, the question "Is it possible to resolve in a general manner the equations whose degree exceeds four?" has yet to receive a satisfactory answer. It is the purpose of this memoir to respond to this question.

The algebraic resolution of an expression is an expression of its roots as algebraic functions of the coefficients. It is therefore necessary to first consider the general form of algebraic functions, and then to see whether it is possible to satisfy the given equation by replacing the unknown with an algebraic function.

III. On the number of distinct values that a function of several variables can assume when these variables are interchanged amongst themselves.

Let v be a rational function of several independent variables x_1, x_2, \dots, x_n . The number of different values to which this function is subjected upon interchanging the quantities on which it depends cannot exceed the product $1 \cdot 2 \cdot 3 \cdots n$. Let μ be this product.

Now let

$$v \begin{pmatrix} \alpha & \beta & \gamma & \delta & \dots \\ a & b & c & d & \dots \end{pmatrix}$$

be the value that some function v assumes when $x_a, x_b, x_c, x_d, \dots$ are substituted for $x_\alpha, x_\beta, x_\gamma, x_\delta, \dots$. It is clear that when A_1, A_2, \dots, A_n denote the μ permutations that can be formed with the indices $1, 2, 3, \dots, n$, the different values of v can be expressed as

$$v \begin{pmatrix} A_1 \\ A_1 \end{pmatrix}, v \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}, v \begin{pmatrix} A_1 \\ A_3 \end{pmatrix}, \dots, v \begin{pmatrix} A_1 \\ A_\mu \end{pmatrix}.$$

¹ *Journal für die reine und angewandte Mathematik* (Crelle), Vol 1, Berlin, 1826

Suppose that the number of distinct values of v is less than μ , it is then necessary that some of the values of v be equal to each other, say

$$v \begin{pmatrix} A_1 \\ A_1 \end{pmatrix} = v \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} = v \begin{pmatrix} A_1 \\ A_1 \end{pmatrix} = \cdots = v \begin{pmatrix} A_1 \\ A_m \end{pmatrix}.$$

If the permutation denoted by

$$\begin{pmatrix} A_1 \\ A_{m+1} \end{pmatrix}$$

is applied to these quantities we will have this new series of equal values:

$$v \begin{pmatrix} A_1 \\ A_{m+1} \end{pmatrix} = v \begin{pmatrix} A_1 \\ A_{m+2} \end{pmatrix} = v \begin{pmatrix} A_1 \\ A_{m+3} \end{pmatrix} = \cdots = v \begin{pmatrix} A_1 \\ A_{2m} \end{pmatrix},$$

values which are different from the first ones, but the same in number. By changing these quantities by the substitution denoted by

$$\begin{pmatrix} A_1 \\ A_{2m+1} \end{pmatrix},$$

we obtain a new system of equal quantities which are, however, different from the preceding ones. By continuing this process until all the permutations have been exhausted, the μ values of v will be partitioned into several systems, each of which will contain m equal values. It follows from this that if the number of distinct values of v is represented by ρ , a number that equals that of the systems, we have $\rho m = 1 \cdot 2 \cdot 3 \cdots n$. That is, the number of distinct values that a function of n quantities can assume under all the possible permutations of these quantities is necessarily a divisor of the product $1 \cdot 2 \cdot 3 \cdots n$. This is known.

Now, let

$$\begin{pmatrix} A_1 \\ A_m \end{pmatrix}$$

be an arbitrary substitution. Suppose that in applying it several times successively to the function v we obtain the sequence of values v, v_1, v_2, \dots, v_p . It is clear that v will be necessarily repeated several times. When v returns after p substitutions we say that

$$\begin{pmatrix} A_1 \\ A_m \end{pmatrix}$$

is a *recurring substitution of order p* . We then have the following periodic series:

$$v, v_1, v_2, \dots, v_{p-1}, v, v_1, v_2, \dots$$

wherein if

$$v \left(\begin{array}{c} A_1 \\ A_m \end{array} \right)^r$$

represents the value of v which is obtained after r repetitions of the substitution denoted by

$$\left(\begin{array}{c} A_1 \\ A_m \end{array} \right),$$

we obtain the series

$$v \left(\begin{array}{c} A_1 \\ A_m \end{array} \right)^0, v \left(\begin{array}{c} A_1 \\ A_m \end{array} \right)^1, v \left(\begin{array}{c} A_1 \\ A_m \end{array} \right)^2, \dots, v \left(\begin{array}{c} A_1 \\ A_m \end{array} \right)^{p-1}, v \left(\begin{array}{c} A_1 \\ A_m \end{array} \right)^0, \dots$$

It follows that

$$v \left(\begin{array}{c} A_1 \\ A_m \end{array} \right)^{\alpha p + r} = v \left(\begin{array}{c} A_1 \\ A_m \end{array} \right)^r \quad \text{and} \quad v \left(\begin{array}{c} A_1 \\ A_m \end{array} \right)^{\alpha p} = v \left(\begin{array}{c} A_1 \\ A_m \end{array} \right)^0 = v.$$

However, if p is the largest prime no greater than n , then if the number of distinct values of v is less than p , it must be the case that amongst p values some two must equal each other.

It therefore follows that of the p values

$$v \left(\begin{array}{c} A_1 \\ A_m \end{array} \right)^0, v \left(\begin{array}{c} A_1 \\ A_m \end{array} \right)^1, v \left(\begin{array}{c} A_1 \\ A_m \end{array} \right)^2, \dots, v \left(\begin{array}{c} A_1 \\ A_m \end{array} \right)^{p-1},$$

some two must equal each other. Say

$$v \left(\begin{array}{c} A_1 \\ A_m \end{array} \right)^r = v \left(\begin{array}{c} A_1 \\ A_m \end{array} \right)^{r'}.$$

Then

$$v \left(\begin{array}{c} A_1 \\ A_m \end{array} \right)^{r+p-r} = v \left(\begin{array}{c} A_1 \\ A_m \end{array} \right)^{r'+p-r}.$$

Writing r for $r' + p - r$ and noting that

$$v \left(\begin{array}{c} A_1 \\ A_m \end{array} \right)^p = v,$$

we conclude that

$$v = v \left(\begin{array}{c} A_1 \\ A_m \end{array} \right)^r$$

where r is clearly not a multiple of p . The value of v is therefore not changed by the substitution

$$\left(\begin{array}{c} A_1 \\ A_m \end{array} \right)^r,$$

nor, consequently, by the repetition of this substitution. We therefore have

$$v = v \left(\begin{array}{c} A_1 \\ A_m \end{array} \right)^{r\alpha}$$

where α is an integer. If p is a prime number, then it is clearly always possible to find two integers α and β such that $r\alpha = p\beta + 1$, hence

$$v = v \left(\begin{array}{c} A_1 \\ A_m \end{array} \right)^{p\beta+1}.$$

Since

$$v = v \left(\begin{array}{c} A_1 \\ A_m \end{array} \right)^{p\beta},$$

it follows that

$$v = v \left(\begin{array}{c} A_1 \\ A_m \end{array} \right).$$

The value of v will therefore not be changed by the recurrent substitution

$$\left(\begin{array}{c} A_1 \\ A_m \end{array} \right)$$

of order p .

However, it is clear that

$$\left(\begin{array}{cccccc} \alpha & \beta & \gamma & \delta & \dots & \zeta & \eta \\ \beta & \gamma & \delta & \varepsilon & \dots & \eta & \alpha \end{array} \right) \quad \text{and} \quad \left(\begin{array}{cccccc} \beta & \gamma & \delta & \varepsilon & \dots & \eta & \alpha \\ \gamma & \alpha & \beta & \delta & \dots & \zeta & \eta \end{array} \right)$$

are recurrent substitutions of order p when p is the number of indices $\alpha, \beta, \dots, \eta$. The value of v will therefore not be changed by the combination of these two. These substitutions are clearly equivalent to the single one

$$\begin{pmatrix} \alpha & \beta & \gamma \\ \gamma & \alpha & \beta \end{pmatrix}$$

and this one to the following two, applied successively:

$$\begin{pmatrix} \alpha & \beta \\ \beta & \alpha \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \beta & \gamma \\ \gamma & \beta \end{pmatrix}.$$

The value of v will therefore not be changed by the combination of these two substitutions. Hence

$$v = v \begin{pmatrix} \alpha & \beta \\ \beta & \alpha \end{pmatrix} \begin{pmatrix} \beta & \gamma \\ \gamma & \beta \end{pmatrix};$$

and similarly

$$v = v \begin{pmatrix} \beta & \gamma \\ \gamma & \beta \end{pmatrix} \begin{pmatrix} \gamma & \delta \\ \delta & \gamma \end{pmatrix},$$

from which it follows that

$$v = v \begin{pmatrix} \alpha & \beta \\ \beta & \alpha \end{pmatrix} \begin{pmatrix} \gamma & \delta \\ \delta & \gamma \end{pmatrix}.$$

We see from this that the function v is unchanged by two successive substitutions of the form $\begin{pmatrix} \gamma & \delta \\ \delta & \gamma \end{pmatrix}$, α and β being any two indices. If such a substitution is called a *transposition*, it may be concluded that any value of v will not be changed by an even number of transpositions, and that consequently all the values of v which result from an even number of substitutions are equal. Every exchange of the elements of a function can be effected by means of a certain number of transpositions; hence the function v can have only two values. The following theorem follows from this: The number of different values that a function of n quantities can assume is not less than the largest prime not exceeding n , unless it is either 2 or 1.

It is therefore impossible to find a function of five quantities which has three or four different values. The demonstration of this theorem is taken from a memoir of Mr. Cauchy which appears in the 17th volume of the *Journal de l'école polytechnique*, p. 1.

D. Excerpts from Galois's *On the Theory of Numbers*¹

When it is agreed to consider as zero all the quantities which are the multiples of a given prime number p , and, subject to this convention, one looks for solutions to the polynomial equation $Fx = 0$, i.e., the equation that Mr. Gauss denotes by $Fx \equiv 0$, it is customary to consider only integer solutions to these sorts of questions. Having been led by some specific researches to consider their irrational solutions, I have arrived at some results that I believe to be new.

Let there be given such an equation or congruence, $Fx = 0$, and let p be the modulus. Suppose first that the congruence in question admits no rational factors, that is, there exist no three polynomials φx , ψx , χx such that $\varphi x \cdot \chi x = Fx + p \cdot \chi x$. In that case the congruence has no integer roots, nor any irrational root of smaller degree. One should therefore regard the roots of this congruence as some kind of imaginary symbols (since they do not satisfy the same questions as integers), symbols whose employment, in calculations, will often prove as useful as that of the imaginary $\sqrt{-1}$ in ordinary analysis.

We are concerned here with the classification of these imaginaries and their reduction to the smallest possible number.

Let i denote one of the roots of the congruence $Fx = 0$, which can be supposed to have degree ν . Consider the general expression

$$a + a_1 i + a_2 i^2 + \cdots + a_{\nu-1} i^{\nu-1} \quad (\text{A})$$

where $a, a_1, a_2, \dots, a_{\nu-1}$ represent integers. When these numbers are assigned all their possible values, expression (A) runs through p^ν values which possess, as I shall demonstrate, the same properties as the natural numbers in the *theory of residues of powers*.

Of the expressions (A), we shall only take the $p^\nu - 1$ values obtained when

$$a, a_1, a_2, \dots, a_{\nu-1}$$

¹ *Bulletin des Sciences mathématiques* de M. Ferussac, Vol 13, Jun 1830, 428–435; with the following note: “This memoir forms part of the research of Mr. Galois on the theory of permutations and algebraic equations.”

are not all zero; let α be one of these expressions.

If α is successively raised to the second, third, ... powers, a sequence of quantities all of which have the same form is obtained (since every function of i is reducible to the $(\nu-1)$ -th degree). Hence it must be that $\alpha^n = 1$ for some n ; let n be the smallest number such that $\alpha^n = 1$. Then the set of numbers $1, \alpha, \dots, \alpha^{n-1}$ are all distinct. Multiply these n numbers by another expression of the same form. We then obtain another new group of quantities all different from the first group as well as from each other. If the quantities (A) have not been exhausted yet, the powers of α can be multiplied by a new expression γ , and so on. Consequently, the number n necessarily divides the total number of quantities of type (A). Since this number is $p^\nu - 1$, we see that n divides $p^\nu - 1$. From this it also follows that $\alpha^{p^\nu-1} = 1$, or $\alpha^{p^\nu} = \alpha$.

Next it can be proven, just as is done in the theory of numbers, that there exist primitive roots α for which $p^\nu - 1 = n$, and which consequently reproduce, by their powers, the complete sequence of all the other roots.

And any one of these primitive roots depends only on a congruence of degree ν , a congruence which must be irreducible, since otherwise the equation of i could not be irreducible either, as the roots of the congruence in i are all powers of the primitive root.

We note here the remarkable result that all the algebraic quantities that arise in this theory are roots of equations of the form $x^{p^\nu} = x$. This proposition is stated algebraically as follows: Given a function Fx and an integer p , one can write $f x \cdot Fx = x^{p^\nu} - x + p \phi x$, $f x$ and ϕx being entire functions, whenever the congruence $Fx = 0 \pmod{p}$ is irreducible.

If it is desired to express all the roots of such a congruence in terms of one, it suffices to note that in general $(Fx)^{p^\nu} = F(x^{p^\nu})$ and that, consequently, if one of the roots is x then the others are $x^p, x^{p^2}, \dots, x^{p^{\nu-1}}$.² We now show that, conversely, the roots of the equation or of the congruence $x^{p^\nu} = x$ all depend on a single congruence of degree ν .

Let i be a root of an irreducible congruence and such that all the roots of the congruence $x^{p^\nu} = x$ are rational functions of i (as is the case for ordinary equations, it is clear that this property holds here as well).³

² It would be wrong to conclude from the fact that the roots of the irreducible congruence $Fx = 0$ of degree ν are expressible as the sequence $x, x^p, x^{p^2}, \dots, x^{p^{\nu-1}}$ that these roots are always expressible by radicals. Here is an example to the contrary: The irreducible congruence $x^2 + x + 1 = 0 \pmod{2}$ yields $x = (-1 + \sqrt{-3})/2$, which reduced to $0/0 \pmod{2}$, from which formula we learn nothing.

³ The general proposition in question here can be stated as follows: Given an algebraic equation, it is possible to find a rational function θ of all of its roots such that, reciprocally, each of the roots is rationally expressible in

It is clear that the degree μ of the congruence in i cannot be less than ν , since otherwise the congruence

$$x^{p^{\mu-1}} = 0 \quad (\text{B})$$

would share all of its roots with the congruence $x^{p^{\mu-1}} = 0$, which is impossible, since the congruence (B) has no repeated roots, as is seen by taking the derivative of the first part. I claim that neither can μ exceed ν .

In fact, if that were the case, all the roots of the congruence $x^{p^\mu} = x$ would depend rationally on those of the congruence $x^{p^\nu} = x$. But it is easily seen that if $i^{p^\nu} = i$, then every rational function $h = f(i)$ would yield

$$(f(i))^{p^\nu} = f(i^{p^\nu}) = f(i)$$

from which $h^{p^\nu} = h$.

Hence all the roots of the congruence $x^{p^\mu} = x$ would also be roots of the equation $x^{p^\nu} = x$, which is impossible.

We therefore now know that all the roots of the equation $x^{p^\nu} = x$ necessarily depend on only one irreducible congruence of degree ν .

Now, the most general method for obtaining this irreducible congruence on which the roots of the congruence $x^{p^\nu} = x$ depend, is to extract first out of this congruence all the factors that it shares with congruences of smaller degree and of the form $x^{p^\mu} = x$.

One thus obtains a congruence which must factor into irreducible congruences of degree ν . And, since it is known how to express all the roots of each of these irreducible congruences in terms of a single one, it will be easy to obtain all of them by Mr. Gauss's method.

Most frequently, however, it will be easy to find by trial and error an irreducible congruence of a given degree ν , and the rest must be derived from it.

For example, let $p = 7$ and $\nu = 3$. Let us look for the roots of the congruence

$$x^{7^3} = x \pmod{7}. \quad (\text{I})$$

θ . This theorem was known to Abel as one can see in the first part of the memoir on elliptic functions which this celebrated geometer left.

I note first that since the congruence

$$i^3 = 2 \pmod{7}, \quad (2)$$

being irreducible, and of degree three, all the roots of congruence (1) depend rationally on those of congruence (2), so that all the roots of (1) have the form

$$a + a_1 i + a_2 i^2 \quad \text{or} \quad a + a_1 \sqrt[3]{2} + s_2 \sqrt[3]{4}. \quad (3)$$

It is now necessary to find a primitive root, that is, a form of expression (3) which, when raised to all possible powers, gives all the roots of the congruence $x^{7^3-1} = 1 \pmod{7}$, or $x^{2 \cdot 3^2 \cdot 19} = 1 \pmod{7}$, and to accomplish this we only need a primitive root of each of the congruences $x^2 = 1$, $x^{3^2} = 1$, and $x^{19} = 1$.

The primitive root of the first is -1 ; those of the second are given by the equations $x^3 = 2$ and $x^3 = 4$, so that i is a primitive root of $x^{3^2} = 1$.

It only remains to find a root of $x^{19} - 1 = 0$, or rather of

$$\frac{x^{19} - 1}{x - 1} = 0,$$

and we first try to see whether the requirements can be satisfied by taking $x = a + a_1 i$ rather than $a + a_1 i + a_2 i^2$; we must have $(a + a_1 i)^{19} = 1$, which, when developed by Newton's formula, after reducing the powers of a , a_1 , and i by applying the formulas $a^{m(p-1)} = 1$, $a_1^{m(p-1)} = 1$, and $i^3 = 2$, reduced to

$$3[a - a^4 a_1^3 + (a^5 a_1^2 + a^2 a_1^5) i^2] = 1,$$

from which, by separation, $3a - 3a^4 a_1^3 = 1$ and $a^5 a_1^2 + a^2 a_1^5 = 0$.

These last two equations are satisfied by setting $a = -1$ and $a_1 = 1$. Hence $-1 + i$ is a primitive root of $x^{19} = 1$. We found above that the values -1 and i are primitive roots of $x^2 = 1$ and $x^{3^2} = 1$; it only remains to multiply the three quantities -1 , i , and $-1 + i$, and the product $i - i^2$ will be a primitive root of the congruence $x^{7^3-1} = 1$.

Thus here the expression $i - i^2$ possesses the property that, in raising it to all powers, $7^3 - 1$ distinct expressions of the form $a + a_1 i + a_2 i^2$ are obtained.

If we wish to find the lowest degree congruence on which our primitive root depends, it is necessary to eliminate i from the two equations $i^3 = 2$ and $\alpha = i - i^2$. One then obtains $\alpha^3 - \alpha + 2 = 0$.

We agree to take imaginaries as a basis and to denote by i the root of this equation, so that

$$i^3 - i + 2 = 0, \quad (C)$$

and we will have all the imaginaries of the form $a + a_1 i + a_2 i^2$ when i is raised to all of its powers and they are reduced by equation (C).

The main advantage of this new theory that is propounded here is that it restores to congruences the property (so useful for ordinary equations) that they possess as many roots as there are units in their degree.

The method of obtaining all of these roots is very simple. First, it is always possible to modify the given congruence $Fx = 0$ so that it does not have equal roots, or, in other words, so that it does not possess a common factor with $F'x = 0$, and the means for doing so are evidently the same as those for ordinary equations.

Next, in order to obtain the integral solutions, it will suffice, as Mr. Libri seems to have been the first to remark, to look for the greatest common factor of $Fx = 0$ and $X^{p-1} = 1$.

To find the imaginary solutions of the second degree, it is necessary to look for the greatest common factor of $Fx = 0$ and $x^{p^2-1} = 1$, and, in general, the solutions of order ν are given by the greatest common factor of $Fx = 0$ and $x^{p^\nu-1} = 1$.

It is above all in the theory of permutations, where it is often necessary to vary the form of the indices, that the consideration of imaginary roots of congruences seems indispensable. It provides a simple and easy method for recognizing in what case a primitive equation is solvable by radicals, as I will attempt to describe in a few words.

Let $fx = 0$ be an algebraic equation of degree p^ν ; suppose that the p^ν roots are denoted by x_k , where the index k assumes the p^ν values determined by the congruence $k^{p^\nu} = k \pmod{p}$.

Let V be an arbitrary rational function of the p^ν roots x_k . Transform this function by replacing each index k by the index $(ak + b)^{p^r}$, a , b , and r being arbitrary constants satisfying the requirements $a^{p^{\nu-1}} = 1 \pmod{p}$, $b^{p^\nu} = b \pmod{p}$ and r integral.

By assigning to the constants a , b , and r all their admissible values, we obtain a total of $p^\nu(p^\nu - 1)\nu$ ways of permuting the roots amongst themselves by means of permutations of the form $[x_k, x_{(ak+b)p^r}]$, and the function V will in general assume $p^\nu(p^\nu - 1)\nu$ different forms as a result of these substitutions.

Assume now that the proposed equation $fx = 0$ is solvable by radicals, and, to prove this result, it suffices to note that the value substituted for k , in each index, can be

expressed in the three forms

$$(ak + b)^{p^r} = [a(k + b^1)]^{p^r} = a'k^{p^r} + b'' = a'(k + b')^{p^r}.$$

Those who are familiar with the theory of equations will see this easily.

This remark would have had little significance had I not succeeded in showing that, conversely, every primitive equation that is solvable by radicals must satisfy the conditions I have just stated. (The equations of the ninth and twenty-fifth degrees are excepted from this rule.)

Thus, for each number of the form p^ν it is possible to form a group of permutations such that every function of the roots that is invariant under the action of these permutations must admit a rational value when the equation of degree p^ν is primitive and solvable by radicals.

Moreover, only equations of such degree p^ν are simultaneously both primitive and solvable by radicals.

The general theorem I have just announced makes precise and develops the conditions that I specified in the *Bulletin* of the month of April. It indicates the means for forming a function of the roots whose value will be rational whenever the primitive equation of degree p^ν is solvable by radicals, and consequently it leads to a characterization of the solvability of these equations by means of calculations which, while perhaps not feasible in practice, are at least theoretically possible.

Note that in the case $\nu = 1$ the various values of k consist of the sequence of integers. There are then $p(p-1)$ substitutions of the form (x_k, x_{ak+b}) .

The function which, in the case of equations that are solvable by radicals, has a rational value depends, in general, on an equation of degree $1 \cdot 2 \cdot 3 \cdots (p-2)$, to which it is necessary, consequently, to apply the method of rational roots.

E. Excerpts from Cayley's *The Theory of Groups*¹

Substitutions, and (in connexion therewith) groups, have been a good deal studied; but only a little has been done towards the solution of the general problem of groups. I give the theory so far as is necessary for the purpose of pointing out what appears to me to be wanting.

Let α, β, \dots be functional symbols, each operating upon one and the same number of letters and producing as its result the same number of functions of these letters; for instance, $\alpha(x, y, z) = (X, Y, Z)$, where the capitals denote each of them a given function of (x, y, z) .

Such symbols are susceptible of repetition and combination; $\alpha^2(x, y, z) = \alpha(X, Y, Z)$, or $\beta\alpha(x, y, z) = \beta(X, Y, Z)$, equal in each case three given functions of (x, y, z) and similarly α^3 , $\alpha^2\beta$, etc.

The symbols are not in general commutative, $\alpha\beta$ not equal to $\beta\alpha$; but they are associative, $\alpha\beta \cdot \gamma = \alpha \cdot \beta\gamma$, each of which is equal to $\alpha\beta\gamma$, which has thus a determinate signification. [The associativeness of such symbols arises from the circumstance that the definitions of α, β, \dots determine the meanings of $\alpha\beta, \alpha\gamma$, etc.: if α, β, \dots were quasi-quantitative symbols such as the quaternion imaginaries i, j , and k , then $\alpha\beta$ and $\beta\gamma$ might have by definition values δ and ε such that $\alpha\beta \cdot \gamma$ and $\alpha \cdot \beta\gamma$, equal to $\delta\gamma$ and $\alpha\varepsilon$, respectively, have unequal values.]

Unity as a functional symbol denotes that the letters are unaltered, $1(x, y, z) = (x, y, z)$, whence $1\alpha = \alpha 1 = \alpha$.

The functional symbols may be substitutions; $\alpha(x, y, z) = (y, z, x)$, the same letters in a different order: substitutions can be represented by the notation $\alpha = \frac{yzx}{xyz}$, the substitution which changes xyz into yzx , or as products of cyclical substitutions, $\alpha = \frac{yzx}{xyz} \frac{uw}{uw} = (x \ y \ z)(u \ w)$, the product of the cyclical interchanges x into y , y into z , and z into x ; and u into w , w into u .

¹ *The American Journal of Mathematics*, 1 (1878), 50–52. Reprinted with permission of The Johns Hopkins University Press.

A set of symbols α, β, \dots such that the product $\alpha\beta$ of each two of them (in each order, $\alpha\beta$ or $\beta\alpha$) is a symbol of the set, is a group. It is easily seen that 1 is a symbol of every group, and we may therefore give the definition in the form that a set of symbols $1, \alpha, \beta, \dots$ satisfying the foregoing condition is a group. When the number of the symbols (or terms) is equal to n , then the group is of the n -th order; and each α is such that $\alpha^n = 1$, so that a group of the order n is, in fact, a group of symbolical n -th roots of unity.

A group is defined by means of the laws of combination of its symbols: for the statement of these we may either (by the introduction of powers and products) diminish as much as may be the number of independent functional symbols, or else, using distinct letters for the several terms of the group, employ a square diagram as presently mentioned.

Thus in the first mode, a group is $1, \beta, \beta^2, \alpha, \alpha\beta, \alpha\beta^2$ ($\alpha^2 = 1, \beta^2 = 1, \alpha\beta = \beta^2 = \alpha$); where observe that these conditions imply also $\alpha\beta^2 = \beta\alpha$. Or, in the second mode calling the same group $(1, \alpha, \beta, \gamma, \delta, \epsilon)$, the laws of combination are given by the square diagram

	1	α	β	γ	δ	ϵ
1	1	α	β	γ	δ	ϵ
α	α	1	γ	β	ϵ	δ
β	β	ϵ	δ	α	1	γ
γ	γ	δ	ϵ	1	α	β
δ	δ	γ	1	ϵ	β	α
ϵ	ϵ	α	ϵ	δ	γ	1

for the symbols $(1, \alpha, \beta, \gamma, \delta, \epsilon)$ are in fact equal to $(1, \alpha, \beta, \alpha\beta, \beta^2, \alpha\beta^2)$.

The general problem is to find all the groups of a given order n ; thus if $n = 2$, the only group is $1, \alpha$ where $\alpha^2 = 1$; $n = 3$, the only group is $1, \alpha, \alpha^2$ where $(\alpha^3 = 1)$; $n = 4$, the groups are $1, \alpha, \alpha^2, \alpha^3$ with $\alpha^4 = 1$ and $1, \alpha, \beta, \alpha\beta$ where $\alpha^2 = 1, \beta^2 = 1$, and $\alpha\beta = \beta\alpha$;² $n = 6$, there are three groups, a group $1, \alpha, \alpha^2, \alpha^3, \alpha^4, \alpha^5$ where $\alpha^6 = 1$; and two groups $1, \beta, \beta^2, \alpha, \alpha\beta, \alpha\beta^2$ where $\alpha^2 = 1$ and $\beta^3 = 1$, viz. in the first of these $\alpha\beta = \beta\alpha$; while in the other of them (that mentioned above) we have $\alpha\beta = \beta^2\alpha$ and $\alpha\beta^2 = \beta\alpha$.

² If $n = 5$, the only group is $1, \alpha, \alpha^2, \alpha^3, \alpha^4$ where $\alpha^5 = 1$.

But although the theory as above stated is a general one, including as a particular case the theory of substitutions, yet the general problem of finding all the groups of a given order n is really identical with the apparently less general problem of finding all of the groups of the same order n , which can be formed with the substitutions upon n letters; in fact, referring to the diagram, it appears that $1, \alpha, \beta, \gamma, \delta, \varepsilon$ may be regarded as substitutions performed upon the six letters $1, \alpha, \beta, \gamma, \delta, \varepsilon$, viz., 1 is the substitution unity which leaves the order unaltered, α the substitution which changes $1\alpha\beta\gamma\delta\varepsilon$ into $\alpha 1\gamma\beta\varepsilon\delta$, and so for β, γ, δ , and ε . This, however, does not in any wise show that the best or easiest mode of treating the general problem is thus to regard it as a problem of substitutions: and it seems clear that the better course is to consider the general problem in itself, and to deduce from it the theory of groups of substitutions.

Cambridge, 26th November, 1877.

F. Mathematical Induction

The Principle of Mathematical Induction (PMI) is not a theorem. It is a powerful method for proving theorems. Other such methods are proof by contradiction, argument by symmetry, and the pigeonhole principle. The following is probably the simplest form of this principle:

Principle of Mathematical Induction, Version 1 Suppose that a set S of positive integers has the two properties that $1 \in S$ and if $k \in S$ then $k + 1 \in S$. The S consists of all the positive integers.

This is eminently reasonable. By the first property, $1 \in S$. The second property therefore implies that $2 = 1 + 1 \in S$. Now that $2 \in S$, it follows from the second property that $3 = 2 + 1 \in S$, and similarly $4 = 3 + 1$, $5 = 4 + 1$, and so on, are all in S .

Let us see how this obvious principle can be used to prove a nonobvious fact.

Theorem F.1 If n is any positive integer, then

$$1 + 2 + 3 + \cdots + n = \frac{n(n+1)}{2}. \quad (\text{F.2})$$

Proof. Let S be the set of all the positive integers n for which Equation F.2 is valid. The PMI will be used to demonstrate that S consists of all the positive integers, which is, of course, tantamount to proving the proposition.

It must first be shown that $1 \in S$. In other words, it must first be shown that when n is replaced by the integer 1 in Equation F.2, a true statement is obtained. This, however, is easily verified, since this replacement transforms Equation F.2 to the statement $1 = 1(1+1)/2$, which is true.

Second, it must be shown that the assumption $k \in S$ leads to the conclusion $k + 1 \in S$. That $k \in S$ means that Equation F.2 is valid when n is replaced by k , so that

$$1 + 2 + 3 + \cdots + k = \frac{k(k+1)}{2}. \quad (\text{F.3})$$

To conclude that $k + 1 \in S$, it must be shown that

$$1 + 2 + 3 + \cdots + k + (k + 1) = \frac{(k + 1)[(k + 1) + 1]}{2}.$$

This is demonstrated, using Equation F.3, as follows:

$$\begin{aligned} 1 + 2 + 3 + \cdots + k + (k + 1) &= \frac{k(k + 1)}{2} + (k + 1) \\ &= \frac{k(k + 1) + 2(k + 1)}{2} \\ &= \frac{(k + 1)(k + 2)}{2} \\ &= \frac{(k + 1)[(k + 1) + 1]}{2}. \end{aligned}$$

Thus the set S enjoys both of the properties required by the PMI, and so it consists of all the positive integers. In other words, Equation F.2 is valid for all the positive integers. ■

The same method is now used to prove a well-known formula that is often demonstrated by other means.

Theorem F.4 If r is any number distinct from 1 and n is any positive integer, then

$$1 + r + r^2 + \cdots + r^n = \frac{1 - r^{n+1}}{1 - r}. \quad (\text{F.5})$$

Proof. Let S be the set of positive integers for which Equation F.5 is valid. Our strategy will be to use the PMI to show that the set S consists of all the positive integers.

For $n = 1$, Equation F.5 reduced to $1 + r = (1 - r^2)/(1 - r)$, which is true, since $1 - r^2 = (1 - r)(1 + r)$. Thus $1 \in S$.

Next, suppose that $k \in S$. In other words, suppose that

$$1 + r + r^2 + \cdots + r^k = \frac{1 - r^{k+1}}{1 - r}. \quad (\text{F.6})$$

Making use of Equation F.6,

$$\begin{aligned}
 1 + r + r^2 + \cdots + r^k + r^{k+1} &= \frac{1 - r^{k+1}}{1 - r} + r^{k+1} \\
 &= \frac{1 - r^{k+1} + (1 - r)r^{k+1}}{1 - r} \\
 &= \frac{1 - r^{k+1} + r^{k+1} - r^{k+1}}{1 - r} \\
 &= \frac{1 - r^{k+1}}{1 - r} = \frac{1 - r^{(k+1)+1}}{1 - r},
 \end{aligned}$$

which means that Equation F.5 holds for $n = k + 1$ as well. In other words, $k + 1 \in S$.

Thus the set S enjoys both of the properties required by the PMI, and so it consists of all the positive integers. In other words, Equation F.5 is valid for all the positive integers. ■

Since the statement to be proved may involve several variables, it is advisable to begin a proof by mathematical induction by stating explicitly to which variable the process is applied. In any such proof, the verification that $1 \in S$ is referred to as *anchoring the induction*, the assumption that $k \in S$ is called the *induction hypothesis*, and the part of the proof that uses the induction hypothesis to prove that $k + 1 \in S$ is the *induction step*.

In all proofs by mathematical induction, the set S consists of the set of all the integers for which a certain statement is true, and it is customary to leave the set S implicit and to speak simply of the set of integers for which the statement holds. With this convention in mind, the PMI can be restated in the following manner:

Principle of Mathematical Induction, Version 2 Suppose that a statement about positive integers possesses the two properties that the statement is true for 1 and if the statement holds for some integer $k \geq 1$, then it also holds for $k + 1$. Then the statement in question holds for all the positive integers.

It is this new version that is employed in the next example.

Theorem F.7 If n is any positive integer, then

$$1 + 3 + 5 + \cdots + (2n - 1) = n^2. \quad (\text{F.8})$$

Proof. We proceed by induction on n . When n is replaced by 1, Equation F.8 becomes $1 = 1^2$, which is valid. Thus the induction process has been anchored.

Next suppose that Equation F.8 is valid for the integer k . In other words, the induction hypothesis is

$$1 + 3 + 5 + \cdots + (2k - 1) = k^2. \quad (\text{F.9})$$

Then

$$\begin{aligned} 1 + 3 + 5 + \cdots + (2k - 1) + [2(k + 1) - 1] &= k^2 + [2(k + 1) - 1] \\ &= k^2 + 2k + 2 - 1 \\ &= k^2 + 2k + 1 \\ &= (k + 1)^2, \end{aligned}$$

and so Equation F.8 also holds for $k + 1$. Thus Equation F.8 possesses the second property. It follows from version 2 of the PMI that Equation F.8 holds for all positive integers. ■

The choice of 1 as the anchoring point for the PMI is merely a convention. Sometimes it is necessary to choose a different starting point, in which case this principle assumes a slightly different form.

Principle of Mathematical Induction, Version 3 Suppose that a statement about integers possesses the following two properties: The statement is true for the integer a . If the statement holds for some integer $k \geq a$, then it also holds for $k + 1$. Then the statement in question holds for all integers that are greater than or equal to a .

Theorem F.10 If $n \geq 7$, then $(3/2)^n > 2n + 1$.

Proof. The statement of this proposition makes it clear that the induction should be anchored at $n = 7$. For this value of n , the inequality is $(3/2)^7 \approx 17.08 < 2 \cdot 7 + 1$, which is true. Next, assume that the inequality is valid for some integer $k \geq 7$. In other words, assume that k is an integer such that $(3/2)^k > 2k + 1$ and $k \geq 7$. Then

$$\left(\frac{3}{2}\right)^{k+1} = \left(\frac{3}{2}\right)\left(\frac{3}{2}\right)^k > \frac{3}{2}(2k + 1) = 3k + \frac{3}{2} = 2k + 3 + k - \frac{3}{2} > 2(k + 1) + 1.$$

Notice that in verifying these steps, it was necessary to make use of both the induction hypothesis $(3/2)^k > 2k + 1$ and the assumption that $k \geq 7$. By version 3 of the PMI, the inequality holds for all $n \geq 7$. ■

The above examples are all algebraic in nature, but the following examples come from calculus and geometry.

Theorem F.11 If $f(x)$ is any differentiable function of x , let $f'(x)$ denote its derivative with respect to x . Using that $(x)' = 1$ and the product rule,

$$[u(x)v(x)]' = u'(x)v(x) + u(x)v'(x),$$

$(x^n)' = nx^{n-1}$ for every positive integer n .

Proof. We proceed by induction on n . When $n = 1$, $(x^1)' = 1$ is given, and so the induction is anchored at 1. Assume next that $(x^n)' = nx^{n-1}$ is true for $n = k$. Then

$$(x^{k+1})' = [x(x^k)]' = (x)'(x^k) + x(x^k)' = 1(x^k) + x(kx^{k-1}) = x^k + kx^k = (k+1)x^k.$$

By the PMI, $(x^n)' = nx^{n-1}$ for all positive integers n . ■

Theorem F.12 In a plane, any n straight lines, no two of which are parallel and no three of which are concurrent, divide the plane into $(n^2 + n + 2)/2$ regions.

Proof. We proceed by induction on n . When $n = 1$, $(n^2 + n + 2)/2 = 2$, and it is clear that any one line divides the plane into two regions. Thus the induction is anchored at $n = 1$.

Assume that this proposition holds for $n = k$, and suppose that we are given $k + 1$ straight lines in a plane, no two of the lines being parallel and no three of them concurrent. Let one of these lines be labeled as q , and suppose that it is temporarily deleted. By the induction hypothesis, the remaining k lines divide the plane into $(k^2 + k + 2)/2$ regions. Restore the line q to its old position. Since q is not parallel to any of the other lines, and since it does not pass through any of their intersections, it follows that those lines cut q into $k + 1$ sections (two of which happen to be infinite). Each of these sections divides one of the old regions into two new regions. This means that the restoration of q raises the region count by $2(k + 1) - (k + 1) = k + 1$. Consequently, the number of

regions determined by the given $k + 1$ lines is

$$\begin{aligned}\frac{k^2 + k + 2}{2} + k + 1 &= \frac{k^2 + k + 2 + 2k + 2}{2} \\ &= \frac{(k^2 + 2k + 1) + (k + 1) + 2}{2} \\ &= \frac{(k + 1)^2 + (k + 1) + 2}{2}.\end{aligned}$$

Thus, by the PMI, we are done. ■

Sometimes, information about k does not transfer to information about $k + 1$. For example, the prime factorization of k sheds no light whatsoever upon the prime factorization of $k + 1$. For such cases, we have yet another form of mathematical induction.

Principle of Mathematical Induction, Version 4 Suppose that a statement about integers possesses the following two properties: The statement is true for an integer a . For any integer $k \geq a$, if the statement is true for $a, a + 1, \dots, k - 1$, then it is also true for k . Then the statement in question holds for all integers that are greater than or equal to a .

Theorem F.13 Every positive integer $n \geq 2$ can be expressed as the product of primes.

Proof. We proceed by induction on n . Since 2 is a prime, $2 = 2$ anchors the induction at $n = 2$. Next, let $k \geq 2$ be a positive integer such that each of the integers $2, 3, \dots, k - 1$ can be factored into a product of primes. If k is prime, then $k = k$ is the required expression. If k is not a prime, then $k = ab$ for some integers $a, b \geq 2$. In that case we also have $a, b \leq k - 1$, so, by the induction hypothesis, both a and b are expressible as products of primes: $a = p_1 p_2 \cdots p_r$ and $b = q_1 q_2 \cdots q_s$. Then $k = ab = p_1 p_2 \cdots p_r q_1 q_2 \cdots q_s$ is the required expression of k as a product of primes. Hence, by version 4 of the PMI, each of the integers $n \geq 2$ is expressible as the product of primes. ■

Principle of Mathematical Induction, Version 5 Every set of positive integers has a least element.

Proof. We assume that version 4 of the PMI holds and suppose, by way of contradiction, that S is a set of positive integers that does not have a least element. For each positive integer n let $P(n)$ be the statement $n \notin S$. Then $P(1)$ is true because otherwise $1 \in S$ and hence 1 would be a minimum of S , contradicting the assumption that S has no minimum. Suppose $P(j)$ is true for $1 \leq j \leq k$, or, in other words, for all $j \in \{1, 2, \dots, k\}$, $j \notin S$.

Now if $k + 1 \in S$ it follows that $k + 1$ would then be the minimal element of S . Hence $k + 1 \notin S$ and it follows that $P(k + 1)$. Thus we have proved that

- $P(1)$ holds; and
- for all $j \in \{1, 2, \dots, k\}$, $P(j) \Rightarrow P(k + 1)$.

Since we are assuming version 4 of the PMI, it follows that $P(n)$ is true for all positive integers n . This, however, implies that S is the empty set, contradicting the hypothesis that it is nonempty. Consequently S must have a minimal element. ■

Exercises F.1

Each of the examples in Exercises F.I.1 to F.I.17 is to be proved by mathematical induction on the integer n .

1. $1^2 + 2^2 + 3^2 + \dots + n^2 = n(n+1)(2n+1)/6$.
2. $1^3 + 2^3 + 3^3 + \dots + n^3 = n^2(n+1)^2/4$.
3. $1 \cdot 2 + 2 \cdot 3 + 3 \cdot 4 + \dots + n(n+1) = n(n+1)(n+2)/3$.
4. $\frac{1}{1 \cdot 2} + \frac{1}{2 \cdot 3} + \frac{1}{3 \cdot 4} + \dots + \frac{1}{n(n+1)} = \frac{n}{n+1}$.
5. $1 \cdot 2 \cdot 3 + 2 \cdot 3 \cdot 4 + 3 \cdot 4 \cdot 5 + \dots + n(n+1)(n+2) = n(n+1)(n+2)(n+3)/3$.
6. $n^2 > 5n + 17$ for $n \geq 8$.
7. $2^n > n^2$ for $n \geq 5$.
8. $3^n > n^3$ for $n \geq 4$.
9. $(4/3)^n > 1 + n$ for $n \geq 8$.
10. $\lim_{x \rightarrow \infty} x^n e^{-x} = 0$ for $n \geq 0$.
11. $\int_0^\infty x^n e^{-x} dx = n!$ for $n \geq 0$.
12. $n^3 + (n+1)^3 + (n+2)^2$ is divisible by 9 for all $n \geq 1$.
13. $\cos \alpha \cos 2\alpha \cos 4\alpha \dots \cos 2^n \alpha = (\sin(2^{n+1} \alpha)) / (2^{n+1} \sin \alpha)$ for $n \geq 0$.
14. The integer $11^{n+2} + 12^{2n+1}$ is divisible by 133 for all $n \geq 0$.
15. Given n planes, no two of which are parallel and no three of which contain the same straight line, they divide space into $(n^3 + 5n + 6)/6$ portions.
16. Every integer of the form $4n + 3$ with $n \geq 0$ has a prime divisor of the same form.

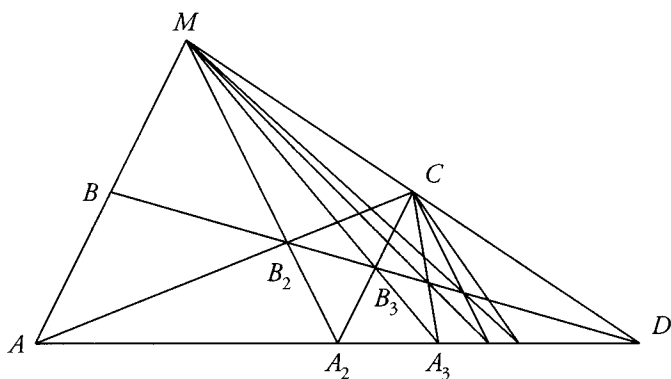


Figure F.1 A geometrical division method

17. Let $ABCD$ be a trapezoid in which the nonparallel sides AB and CD intersect in a point M (see Figure F.1). Define $A_1 = A$, $B_1 = B$, and for each positive integer k , let B_{k+1} be the intersection of the lines CA_k and BD , and let A_{k+1} be the intersection of the lines MB_{k+1} and AD . Then $DA_n = DA/n$ for all $n \geq 1$.
18. Let $a_0 = a_1 = 1$ and $a_{n+2} = a_{n+1} + 5a_n$ for $n \geq 0$. Prove that $a_n \leq 3^n$ for all $n \geq 3$.

G. Logic, Predicates, Sets, and Functions

Propositions are assertions that may be either true or false. If p is a proposition, we write $\lambda(p) = T$ or $\lambda(p) = F$ according as p is true or false.

Propositions can be compounded and *propositional calculus* is concerned with the questions that arise from this compounding. Given two propositions p and q , their *conjunction* is the compound proposition p and q , denoted by $p \wedge q$. For example, if p is the proposition $-1 < 1$ and q is the proposition $-3 < -1$ then we write $p \wedge q$ as either $-1 < 1$ and $-3 < -1$ or, more briefly, as $-3 < -1 < 1$. The *disjunction* of p and q , denoted by $p \vee q$, is the proposition either p or q . Thus, for the same p and q as above, $p \vee q$ is either $-1 < 1$ or $-3 < -1$. Similarly, if p is 10 is divisible by 2 and q is 10 is divisible by 3, then $p \wedge q$ is 10 is divisible by both 2 and 3, and $p \vee q$ is 10 is divisible by either 2 or 3 or both.

G.1 Truth Tables

We stipulate that the logical value of a compound proposition depends only on the logical values of its components. Thus, the conjunction $p \wedge q$ is true if and only if both p and q are true. On the other hand, $p \vee q$ is true if and only if at least one of p and q are true. This is summarized in Table G.1. Such tables are called *truth tables*. Two compound propositions are said to be *equivalent* if they have identical columns in the appropriate truth table. This relationship is denoted by the \equiv symbol. For example, Table G.2 demonstrates that $p \wedge (q \vee r) \equiv (p \wedge q) \vee (p \wedge r)$.

The symbol $\neg p$ denotes the *negation* of p . It has the truth table in Table G.3.

p	q	$p \wedge q$	$p \vee q$
F	F	F	F
F	T	F	T
T	F	F	T
T	T	T	T

Table G.1 Truth tables of conjunction and disjunction

p	q	r	$q \vee r$	$p \wedge (q \vee r)$	$p \wedge q$	$p \wedge r$	$(p \wedge q) \vee (p \wedge r)$
F	F	F	F	F	F	F	F
T	F	F	F	F	F	F	F
F	T	F	T	F	F	F	F
F	F	T	T	F	F	F	F
T	T	F	T	T	T	F	T
T	F	T	T	T	F	T	T
F	T	T	T	F	F	F	F
T	T	T	T	T	T	T	T

Table G.2 The proof of an equivalence

p	$\neg p$
F	T
T	F

Table G.3 The truth table of negation

p	q	$p \Rightarrow q$	$\neg p$	$\neg p \vee q$
F	F	T	T	T
F	T	T	T	T
T	F	F	F	F
T	T	T	F	T

Table G.4 The truth table of implication

Exercises G.1

Prove the equivalences in Exercises G.1.1 to G.1.8 by the construction of appropriate tables.

1. $p \vee q \equiv q \vee p$
2. $p \wedge q \equiv q \wedge p$
3. $(p \vee q) \vee r \equiv p \vee (q \vee r)$
4. $(p \wedge q) \wedge r \equiv p \wedge (q \wedge r)$
5. $(p \vee q) \wedge r \equiv (p \wedge r) \vee (q \wedge r)$
6. $(p \wedge q) \vee r \equiv (p \vee r) \wedge (q \vee r)$
7. $\neg(p \vee q) \equiv \neg p \wedge \neg q$
8. $\neg(p \wedge q) \equiv \neg p \vee \neg q$

G.2 Modeling Implication

It is necessary to define a propositional operator that models implication. This is, after all, a mathematics textbook and implications and consequences are the business of mathematics. If p and q are any two propositions, then we denote the notion that “ p implies q ” by $p \Rightarrow q$. We now argue that the nature of implication is described by the (third column) of Table G.4. The first two rows together assert that a false proposition implies any proposition whatsoever. In other words, a false hypothesis can be used to draw any conclusion whatsoever. This may seem somewhat too sweeping a statement, but consider the following well-known anecdote:

Lord Bertrand Russell, one of the pioneers of mathematical logic, was delivering a talk on this subject, when he was interrupted by someone in the audience who shouted to him “One equals two—prove you are the Pope!” Replied Russell: “The Pope and I are a set of two” To complete Russell’s reasoning, if you assume that one equals two, then the set consisting of him and the Pope is a one-element set from which it follows that the

Pope and Russell are one and the same. Thus, a false statement can imply a completely unrelated false statement.

While it would be all right for a false statement to imply a false one (first row), can a false statement imply a true one (second row)? The answer is yes as shown by the following argument wherein an obviously false statement leads to an equally obvious true statement:

$$-1 = 1 \Rightarrow (-1)^2 = (1)^2 \Rightarrow 1 = 1.$$

Consequently the first two entries in the third column of Table G.3, the ones that correspond to $\lambda(p) = F, \lambda(q) = F$ and $\lambda(p) = F, \lambda(q) = T$, must be T.

On the other hand, the derivation of a false statement from a valid one can never be tolerated under any circumstances whatsoever. It is inherent in human reasoning that any derivation of a false statement, such as $-1 = 1$, from valid hypotheses must contain an error. Hence the entry that corresponds to $\lambda(p) = T, \lambda(q) = F$ must itself be false. This leaves the row corresponding to $\lambda(p) = T, \lambda(q) = T$. The entry must be the same regardless of the instances of p and q , as long as they are both true. The entry under $p \Rightarrow q$ is clearly T when p and q are identical propositions and so it must be T whenever p and q have the same truth value.

The theorems, propositions, lemmas, and whatnots of mathematics are implications, i.e., have the form $p \Rightarrow q$. When simplifying a complex expression of this type it usually comes in handy to massage the complicated expression until it is simplified. Rigor, however, must be preserved.

The *converse* of the statement $p \Rightarrow q$ is the implication $q \Rightarrow p$. For example, if p is $x = -1$ and q is $x^2 = 1$, then it is clear that $p \Rightarrow q$. The converse of this implication is $q \Rightarrow p$ or if $x^2 = 1$ then $x = 1$, which is clearly false. Hence the converse implication $q \Rightarrow p$ is not necessarily valid. The implication $\neg p \Rightarrow \neg q$ is the *inverse* of $p \Rightarrow q$ which is also not logically equivalent since it asserts that if $x \neq 1$ then $x^2 \neq 1$. This is evidently invalid as demonstrated by using $x = -1$.

Thus, so far each time we messed with the aspects of the implication the logical value also was damaged. However, we do hit the jackpot for the *contrapositive*, namely the statement $\neg q \Rightarrow \neg p$. If we proceed, as before, to substitute the same $x = -1$ and $x^2 = 1$, then this contrapositive assumes the form $x^2 \neq 1 \Rightarrow x \neq -1$. We are now faced with the question of the validity of this implication. Contrary to the cases of the converse and the inverse, checking with $x = -1$ doesn't lead to trouble and so the possibility that the contrapositive is valid is still open. We could try other values for x but the results would

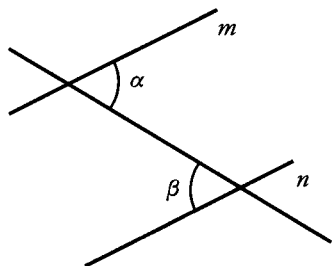


Figure G.1 Euclid's Axiom

	identity	converse	inverse	contrapositive
identity	identity	converse	inverse	contrapositive
converse	converse	identity	contrapositive	inverse
inverse	inverse	contrapositive	identity	converse
contrapositive	contrapositive	inverse	converse	identity

Table G.5 Interrelationships among contrapositive, converse, and inverse

be the same. So, there is hope that the contrapositive is logically equivalent to the original statement. In fact this can be easily proved by constructing the appropriate Truth Table.

The reader might well wonder at all the bruhaha about the contrapositive. After all, all we are saying is that two implications are logically equivalent. However, psychologically speaking one of the implications may be more pliant or easy than the other and hence preferable. For example the implication $x^2 \neq 1 \Rightarrow x \neq -1$ is clearly more complicated and opaque than its contrapositive $x = 1 \Rightarrow x^2 = 1$. Another example is provided by Euclid's Axiom which states that if in Figure G.1 $m \parallel n \Rightarrow \alpha = \beta$ and its contrapositive is $\alpha \neq \beta \Rightarrow m \cap n \neq \emptyset$.

The notions of contrapositive, converse, and inverse have interesting interrelationships. To begin with, the successive application of two distinct notions results in the third one. For example, the converse of the inverse of $p \Rightarrow q$ is the converse of $\neg p \Rightarrow \neg q$ which is $\neg q \Rightarrow \neg p$, which is the contrapositive of $p \Rightarrow q$.

In addition, it is clear that the combined effect of the application of two identical notions results in no change. These observations are summarized in the diagram of Table G.5. The reader may notice that this table is isomorphic to the multiplication table of the Klein 4-group.

Exercises G.2

1. Prove that $p \Rightarrow q \equiv \neg q \Rightarrow \neg p$.
2. Prove that $p \Rightarrow q \not\equiv q \Rightarrow p$.
3. Prove that $p \Rightarrow q \not\equiv \neg p \Rightarrow \neg q$.

G.3 Predicates and Their Negation

We now set out to explain the logic that underlies the intricate use of negation with which the proof of Proposition 11.5 begins. To begin with, we assume that the universe, or domain of possible values for each of our variables, is \mathbb{R} . Let $\varphi(x)$ be a statement about the variable x that becomes either true or false when the variable x is replaced by a specific number. For example $\varphi(x)$ can be any of the following: $\varphi_1(x) : x < 0$, $\varphi_2(x) : x^2 + 1 = 0$, or $\varphi_3(x) : x^2 - 1 = (x - 1)(x + 1)$, where “ $\varphi(x) : \alpha$ ” means that $\varphi(x)$ is α .

Note that $\varphi_1(x)$ is true for some x 's and false for others. On the other hand $\varphi_2(x)$ is false for all x 's, whereas $\varphi_3(x)$ is true for all x 's. Let $\neg\varphi(x)$ denote the negation of $\varphi(x)$. Thus, $\neg\varphi_1(x) : x \geq 0$, $\neg\varphi_2(x) : x^2 + 1 \neq 0$, and $\neg\varphi_3(x) : x^2 - 1 \neq (x - 1)(x + 1)$.

In general, the statement $\varphi(x)$ may be neither true nor false. As noted above, one way to turn it into a proposition that may be true or false is to substitute a real number for x . Thus, $\varphi_1(-1) : -1 < 0$ is true; $\varphi_1(1) : 1 < 0$ is false; $\varphi_2(1) : 1^2 + 1 = 0$ is false; and $\varphi_3(-1) : (-1)^2 - 1 = (-1 - 1)(-1 + 1)$ is true.

Another way to turn the neutral statement $\varphi(x)$ into a proposition that is either true or false is via the application of quantifiers. Let $\exists x$ denote the phrase “there exists an x such that” and let $\forall x$ denote the phrase “for all x .” Then

$$\exists x \varphi_1(x) : \text{there exists a number } x \text{ such that } x < 0,$$

which is true, and

$$\forall x \varphi_1(x) : \text{every number } x \text{ is negative,}$$

which is false. The quantifier \exists is known as the *existential quantifier* since it asserts the existence of some object in the universe. The quantifier \forall is the *universal quantifier* because it asserts that something is true for all the objects. The negation of quantified propositions follows the rules $\neg \exists x \varphi(x) = \forall x \neg \varphi(x)$ and $\neg \forall x \varphi(x) = \exists x \neg \varphi(x)$.

Thus, the negation of the proposition “all numbers are negative,” which is denoted by $\neg \forall x \varphi_1(x)$, is the proposition “there is a number which is not negative,” denoted

by $\exists x \neg \varphi_1(x)$. Similarly, the negation of the proposition “there is a negative number”, denoted by $\neg \exists x \varphi_1(x)$ is the proposition “all numbers are nonnegative,” denoted by $\forall x \neg \varphi_1(x)$.

These considerations become nonobvious when the number of variables is raised. Suppose $\varphi(x, y) : x > y$. We turn this into a proposition by quantifying both of the variables. Say, for starters, that the universal quantifier is applied to both x and y , so that we have the proposition $\forall x \forall y \varphi(x, y)$ whose negation, by two applications of the rules above, is

$$\neg \forall x \forall y \varphi(x, y) = \exists x \neg \forall y \varphi(x, y) = \exists x \exists y \neg \varphi(x, y)$$

where $\neg \varphi(x, y)$ means $x \leq y$. Somewhat less formally, the equation above says that the negation of the proposition $x > y$ for all x and y is $x \leq y$ for some x and some y .

Next let us apply mixed quantifiers to $\varphi(x, y)$, the universal one first, so that we get the proposition $\exists x \forall y \varphi(x, y)$, which is “there exists a number x that is greater than every number y ,” which is clearly false. The negation of this proposition is

$$\neg \exists x \forall y \varphi(x, y) = \forall x \neg \forall y \varphi(x, y) = \forall x \exists y \neg \varphi(x, y),$$

or “for every number x there exists a number y such that $x \leq y$.” This, of course, is a true proposition.

Similarly the proposition $\forall x \exists y \varphi(x, y)$ says “for every x there exists a y such that $x > y$,” which is true. The negation of this proposition is

$$\neg \forall x \exists y \varphi(x, y) = \exists x \neg \exists y \varphi(x, y) = \exists x \forall y \neg \varphi(x, y),$$

which says that “there exists an x such that for all y , $x \leq y$ ” and which is clearly false.

G.4 Two Applications

Let x , a , and L be fixed real numbers and let ε and δ be variables each of whose universes consists of the positive real numbers. Set $p : |x - a| < \delta$ and $q : |f(x) - L| < \varepsilon$. Then the proposition “ $f(x)$ converges to L in the ε - δ sense” is represented by $\forall \varepsilon \exists \delta (p \Rightarrow q)$. Hence when we assume that $f(x)$ fails to converge to L we are assuming the negation $\neg \forall \varepsilon \exists (p \Rightarrow q)$, which is equivalent to

$$\exists \varepsilon \forall \delta \neg (p \Rightarrow q) \equiv \exists \varepsilon \forall \delta \neg (\neg p \vee q) \equiv \exists \varepsilon \forall \delta (p \wedge \neg q).$$

This, however, translates to “there exists a positive ε such that for all positive δ $|x - a| < \delta$ and $|f(x) - L| \geq \varepsilon$.”

Throughout the discussion above it was assumed that the universe of each of the variables x and y is \mathbb{R} . This, of course, need not be the case. Of particular interest is the case where the universe of y is \mathbb{N} , that is, the natural numbers $1, 2, 3, \dots$, whereas ε can be any positive real number. For example, let $\{x_n\}$ be a fixed sequence and define T_m to be the subsequence $T_m = x_{m+1}, x_{m+2}, x_{m+3}, \dots$. Consider the statement $\{x_n\} \rightarrow A$ whose definition is “every neighborhood of A contains all but a finite number of the x_n ’s,” or, equivalently, “for every $\varepsilon > 0$ there exists a natural number m such that $T_m \subseteq (A - \varepsilon, A + \varepsilon)$.” Setting $\varphi(\varepsilon, m) : T_m \subseteq (A - \varepsilon, A + \varepsilon)$ we obtain $\forall \varepsilon \exists m \varphi(\varepsilon, m)$ as the formalization of the convergence $\{x_n\} \rightarrow A$. The negation therefore is

$$\neg \forall \varepsilon \exists m \varphi(\varepsilon, m) = \exists \varepsilon \neg \exists m \varphi(\varepsilon, m) = \exists \varepsilon \forall m \neg \varphi(\varepsilon, m),$$

i.e., “there exists an ε such that for all m , $T_m \not\subseteq (A - \varepsilon, A + \varepsilon)$.”

Exercises G.4

In Exercises G.4.1 to G.4.4, interpret the propositions $\forall x \exists y \varphi(x, y)$, $\exists x \forall y \varphi(x, y)$, $\exists x \exists y \varphi(x, y)$, and $\forall x \forall y \varphi(x, y)$ for the given specification of $\varphi(x, y)$. Decide whether the resulting propositions are true or false.

1. $x < y$ 2. $x = y + 1$ 3. $x = y^2 + 1$ 4. $x^2 + y^2 \leq 1$

Use the results of this section to negate the following propositions.

5. Every person has a progenitor.
6. Every person has a biological child.
7. Every real number has a square root.
8. Every real number has a cubic root.
9. Every rational number has a rational square root.
10. Some real numbers have no real square roots.
11. Some rational numbers have no real square roots.
12. All the integer divisors of 12 are prime.

- 13. None of the people in this room are friendless.
- 14. Every person in this room has a friend.
- 15. There exist two integers whose sum is 5 and whose product is 8.
- 16. The sum of any two integers is smaller than their product.
- 17. Any two cosets are either equal or disjoint.

G.5 Sets

A set is a collection of distinct *elements*. Two sets are *equal* when they have exactly the same elements, regardless of the orders in which their elements are listed. Sets can be denoted in a variety of ways. For example, the set consisting of the integers 2, 3, and 4 is denoted by $\{2, 3, 4\}$. The set consisting of all the squares of positive integers is $\{1, 4, 9, \dots, n^2, \dots\}$ or $\{n^2 \mid n = 1, 2, 3, \dots\}$. The elements of a set have to be distinct from each other and if some repetitions occur they must be dropped. Thus $\{n^2 \mid n = -2, -1, 0, 1, 2\} = \{0, 1, 4\}$.

If the set A contains an element a , we can also say that A *contains* a and we write $a \in A$. If a does not belong to A , this is denoted by $a \notin A$. The *empty set* is the unique set which contains no elements. It is usually denoted by the symbol \emptyset . The meaning of the *universal set*, or the *universe*, depends on the context. If we are working with integers the universal set is \mathbb{Z} . On the other hand, if we are dealing with irrational numbers, the context may be either \mathbb{R} or \mathbb{C} .

If A and B are sets, then $A - B = \{a \in A \mid a \notin B\}$. If A is any set, and the universal set is U , then the *complement* of A (in U) is $A^c = \{a \in U \mid a \notin A\} = U - A$. The *union* of the two sets A and B is $A \cup B = \{c \mid c \in A \text{ or } c \in B \text{ (or both)}\}$ and the *intersection* of A and B is $A \cap B = \{c \mid c \in A \text{ and } c \in B\}$. If every element of A is also an element of B , we write $A \subseteq B$ and say that A is a *subset* of B and that B is a *superset* of A .

The equivalences below connect set theory to the propositional calculus:

$$[x \in (A \cup B)] \equiv (x \in A) \vee (x \in B),$$

$$[x \in (A \cap B)] \equiv (x \in A) \wedge (x \in B),$$

$$[x \in (A - B)] \equiv (x \in A) \wedge (x \notin B).$$

They allow us to derive set theoretic proofs from logical ones. As an example we prove the following theorem.

Theorem G.1 $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

Proof. For any x in our implicit universe

$$\begin{aligned}
 x \in A \cap (B \cup C) &\equiv (x \in A) \wedge (x \in B \vee x \in C) \\
 &\equiv (x \in A \wedge x \in B) \vee (x \in A \wedge x \in C) \\
 &\equiv x \in A \cap B \vee x \in A \cap C \equiv x \in (A \cap B) \cup (A \cap C)
 \end{aligned}$$

and so, by the transitivity of equivalence, $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$. ■

Exercises G.5

Let A, B, C be subsets of the set S . In Exercises G.5.1 to G.5.9, prove the given identity using the method of your choice.

1. $A \cup B = B \cup A$
2. $A \cap B = B \cap A$
3. $A \cup (B \cup C) = (A \cup B) \cup C$
4. $A \cap (B \cap C) = (A \cap B) \cap C$
5. $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
6. $(A \cup B)^c = A^c \cap B^c$
7. $(A \cap B)^c = A^c \cup B^c$
8. $A - (B - C) = (A - B) \cup (A \cap C)$
9. $A - (B \cup C) = (A - B) - C$
10. Let Δ be defined by $A \Delta B = (A - B) \cup (B - A)$. Prove that $A \Delta (B \Delta C) = (A \Delta B) \Delta C$.

G.6 Functions

Given two sets A and B , a function $f : A \rightarrow B$ is a rule that assigns to every element $a \in A$ a unique corresponding element $f(a) \in B$. The *domain* of this function f is A and its *range* is B . The *codomain* of f is the set of all the elements $b \in B$ for which there exists an $a \in A$ such that $f(a) = b$. The codomain of f is denoted by $f(A)$. The two functions $f, g : A \rightarrow B$ are said to be equal provided $f(x) = g(x)$ for all $x \in A$. It is useful to represent functions $f : A \rightarrow B$ by means of a diagram such as Figure G.2.

For any set A we denote by I_A the function $I_A : A \rightarrow A$ such that $I_A(a) = a$ for all $a \in A$. For any sets A, B, C and functions $f : A \rightarrow B$ and $g : B \rightarrow C$, we define the *composition* $g \circ f : A \rightarrow C$ as $(g \circ f)(a) = g(f(a))$ for all $a \in A$.

Proposition G.2 The composition of functions is associative. That is, given functions $f : A \rightarrow B$, $g : B \rightarrow C$, and $h : C \rightarrow D$, then $h \circ (g \circ f) = (h \circ g) \circ f$.

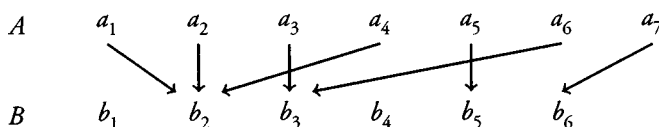


Figure G.2 A diagram of a function

Proof. By several applications of the definition of composition it follows that for any $a \in A$

$$(h \circ (g \circ f))(a) = h((g \circ f)(a)) = h(g(f(a))) = (h \circ g)(f(a)) = ((h \circ g) \circ f)(a).$$

The desired conclusion then follows the definition of equality of functions. ■

Proposition G.3 Let $f : A \rightarrow B$. Then $I_B \circ f = f = f \circ I_A$.

Proof. For any $a \in A$ and $b \in B$

$$(I_B \circ f)(a) = I_B(f(a)) = f(a) = f(I_A)(a) = (f \circ I_A)(a).$$

The definition of the equality of functions now clearly implies the proposition. ■

The function $f : A \rightarrow B$ is said to be *injective* if it has the property that for any $a, a' \in A$ if $a \neq a'$ then $f(a) \neq f(a')$, or, equivalently, if $f(a) = f(a')$ then $a = a'$. In terms of the arrow presentation of functions, f is injective if no distinct arrowtips meet. The function $f : \mathbb{Z} \rightarrow \mathbb{Z}$ defined by $f(m) = m^3$ is injective as are $f_k(m) = m^{2k-1}$ for every positive integer k . On the other hand, the function $f(m) = m^2$ fails to be injective since $f(-3) = 9 = f(3)$.

Proposition G.4 The function $f : A \rightarrow B$ is injective if and only if there exists a function $g : B \rightarrow A$ such that $g \circ f = I_A$.

Any function g that satisfies $g \circ f = I_A$ is said to be a *left inverse* of f . Thus, Proposition G.4 can be rephrased as

Proposition G.5 The function $f : A \rightarrow B$ is injective if and only if it has a left inverse.

Proof. Assume first that $f : A \rightarrow B$ has a left inverse g . If a, b are in A and $f(a) = f(b)$, then

$$a = (g \circ f)(a) = g(f(a)) = g(f(b)) = (g \circ f)(b) = b,$$

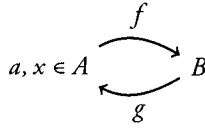


Figure G.3 A left inverse

and hence f is injective.

Conversely, suppose f is injective. We define a function $g : B \rightarrow A$ as follows (see Figure G.3):

$$g(y) = \begin{cases} x & \text{if } f(x) = y; \\ \text{any element of } A & \text{otherwise.} \end{cases}$$

We now demonstrate that g is a left inverse of f . If $a \in A$, then, by the definition of g , $g(f(a)) = a$ and hence $g \circ f$ is the required left inverse. ■

A function $f : A \rightarrow B$ is said to be *surjective* if it has the property that $f(A) = B$, or, in other words, for each $b \in B$ there exists an $a \in A$ such that $f(a) = b$. In terms of the arrow presentation, every element of B is the target of some arrow.

Proposition G.6 The function $f : A \rightarrow B$ is surjective if and only if there exists a function $h : B \rightarrow A$ such that $f \circ h = I_B$.

Any function h that satisfies $f \circ h = I_B$ is said to be a *right inverse* of f . Thus, the proposition above can be rephrased as

Proposition G.7 The function $f : A \rightarrow B$ is surjective if and only if it has a right inverse.

Proof. Assume $f : A \rightarrow B$ is surjective. If $y \in B$ there exists an $x \in A$ such that $f(x) = y$ and we define $h(y)$ to be any x^* such that $h(y) = x^*$. Then

$$(f \circ h)(y) = f(h(y)) = f(x^*) = y = I_B(y)$$

and hence $f \circ h = I_B$.

Conversely, suppose $f : A \rightarrow B$ has the right inverse $h : B \rightarrow A$, so that $f \circ h = I_B$. Then $f(h(b)) = b$ so that f maps $h(b)$ onto the arbitrary b (see Figure G.4). ■

A function is said to be *bijective* if it is both injective and surjective.

Proposition G.8 The function $f : A \rightarrow B$ is bijective if and only if there exists a (necessarily unique) function $g : B \rightarrow A$ such that $g \circ f = I_A$ and $f \circ g = I_B$.

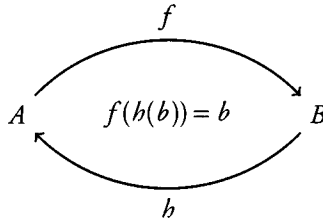


Figure G.4 A right inverse

A bijective function $\sigma : A \rightarrow B$ is called a *permutation*.

Proposition G.9 Let $f : A \rightarrow A$ be A be a finite set. Then the following are equivalent: f is bijective; f is injective; f is surjective.

Proof. See Exercise G.6.15. ■

A *relation* R of the set A is a rule that assigns some elements of A to some elements of A . If the element $b \in A$ is assigned to a , we write $a R b$. A relation R on the set A is an *equivalence relation* if the following three properties hold for any three elements $a, b, c \in A$:

- $a R a$ (Reflexivity);
- $a R b$ implies $b R a$ (Symmetry);
- $a R b$ and $b R c$ imply $a R c$ (Transitivity).

Let R be the relation on the points of the plane defined by $a R b : a = b$. Since for all points a, b , and c , $a = a$, $(a = b) \Rightarrow (b = a)$, and $(a = b) \wedge (b = c) \Rightarrow (a = c)$, it follows that this R (better known as *equality of points*) is an equivalence relation. Similarly, equality of length is an equivalence relation on line segments in either the real line or the plane.

The *raison d'être* of relations is to formalize and clarify the notion of a coset. Let G be a group and H a subgroup of G . Define $a R b$ provided $a^{-1}b \in H$. Now $a R a$ because $a^{-1}a = 1_G \in H$. It is clear that the following are equivalent: $a R b$; $a^{-1}b \in H$; $(a^{-1}b)^{-1} \in H$; $b^{-1}a \in H$; $b R a$. Finally, if $a R b$ and $b R c$, then $a^{-1}b \in H$ and $b^{-1}c \in H$. Multiplication of these equations yields

$$a^{-1}c = a^{-1}b b^{-1}c \in HH = H$$

and so $a R c$.

Proposition G.10 Let R be an equivalence relation on the set A . Then R defines a partition $\mathcal{A} = \{A_\lambda \mid \lambda \in \Lambda\}$ of A such that $a R b$ if and only if a and b belong to the same member of \mathcal{A} .

Proof. For each $a \in A$ let $[a]$ denote the set of all the elements $b \in A$ such that $a R b$. The reflexivity of R implies that for all $a \in A$ $a \in [a]$. The symmetry of R implies the equivalence of the statements $a \in [b]$ and $b \in [a]$. These sets have the property that for each pair $[a]$ and $[b]$ of such sets, either $[a] \cap [b] = \emptyset$ or $[a] = [b]$, for, if $[a] \cap [b] \neq \emptyset$, say, $c \in [a] \cap [b]$, then $c \in [a]$ and $c \in [b]$. Consequently $[a] = [c] = [b]$. Conversely, if a and b belong to the same member of \mathcal{A} , then $[a] = [b]$ and hence $a R b$. ■

Exercises G.6

In Exercises G.6.1 to G.6.12, decide whether the relation R on the set A is an equivalence relation. If yes, describe the equivalence classes. If not, then decide which property is violated.

1. A : people, R : brother of
2. A : people, R : sibling of
3. $A = \{1, 2, \dots, 21\}$, R : $a + b$ is even
4. $A = \{1, 2, \dots, 21\}$, R : $a + b$ is even
5. $A = \{1, 2, \dots, 21\}$, R : $a - b$ is even
6. $A = \{1, 2, \dots, 21\}$, R : $a - b$ is even
7. $A = \{1, 2, \dots, 21\}$, R : $a + b$ is odd
8. $A = \{1, 2, \dots, 21\}$, R : $a + b$ is odd
9. $A = \{1, 2, \dots, 21\}$, R : $|a - b| = 3$
10. $A = \{1, 2, \dots, 21\}$, R : $|a + b| = 2$
11. $A = \mathbb{Z}$, R : $x - y$ is divisible by 2
12. $A = \mathbb{Z}$, R : $x - y$ is divisible by 3
13. Prove that the composition of two injective functions is also injective.
14. Prove that the composition of two surjective functions is also surjective.
15. Prove Proposition G.9.

Biographies

Niels Henrik Abel (1802–1829). Despite his early death from consumption, the Norwegian Abel exerted a major and lasting influence on the evolution of mathematics. He is best known for his work on the quintic equation and elliptic integrals.

Muhammad ibn Musa al-Khwarezmi (ca. 780–850). Al-Khwarezmi was on the faculty of the *House of Wisdom*, a scholarly institute in the city of Baghdad. The author of several mathematical and astronomical works, his name eventually gave rise to the term *algorithm*, and the first two syllables of his text *al-jabr wa'al-muqabala* mutated into the term *algebra*.

Archimedes (287–212 BCE). The Greek Archimedes was the greatest of the scientists, mathematicians, and engineers of antiquity. He lived in the city of Syracuse in Sicily. Among his accomplishments are the formulation of a precise theory of flotation, the computation of the volumes of many solids, including the sphere, and the construction of a variety of engines to defend his city against the besieging Romans.

Rafael Bombelli (1526–1572). An Italian engineer by profession, Bombelli also wrote a widely read treatise called *Algebra*. This book contains the first known attempt to systematize complex numbers.

Gerolamo Cardano (1507–1576). A physician by profession, the Italian Cardano wrote the very influential text *Ars Magna* which included the solutions to the general cubic and quartic equations.

Augustin-Louis Cauchy (1789–1857). The Frenchman Cauchy was the most prolific of all the mathematicians of the nineteenth century. While he worked in many areas of mathematics, he is best remembered today as the founder of the theory of complex variables and also for his contributions to the rigorization of calculus.

Arthur Cayley (1821–1895). Cayley was the most productive English mathematician to follow Newton. He made many contributions to geometry, algebra, and analysis, but is best known for his work on invariant theory.

Richard Dedekind (1831–1916). Dedekind was a German mathematician noted for his contributions to the rigorization of both analysis and algebra. This was facilitated in both cases by the fortuitous idea that missing numbers can be represented by sets of known numbers—Dedekind cuts in analysis and ideals in algebra.

Gotthold Eisenstein (1823–1852). Like his contemporaries Galois and Abel, Eisenstein died young. His death was at least partially due to his one-day arrest by the Prussian army on suspicion of Republicanism. Nevertheless he was prolific (23 papers in 1842 alone) and Gauss thought very highly of his mathematical talents. He contributed much to the development of the new mathematical field of algebraic number theory.

Euclid (3rd century BCE). Euclid was a Greek who lived in Alexandria. He wrote several books of which the best known is *The Elements*. This textbook on geometry and number theory is arguably the most influential scientific tract of all time.

Leonhard Euler (1707–1783). Together with the Frenchman Lagrange, the Swiss Euler completely dominated the mathematical developments of his time. He made fundamental contributions to all areas of mathematics, including some which did not begin to flourish until a century after his death.

Pierre de Fermat (1601–1655). Known as the “Prince of Mathematics,” the Frenchman Fermat was in fact a lawyer who regarded mathematics as a diversion. He did pioneering work in calculus and probability, and set number theory on a course it still follows.

Lodovico Ferrari (1522–1565). A student of Cardano, Ferrari was the first to derive a formula for the solution of the general quartic equation.

Scipione del Ferro (1465–1526). The Italian del Ferro was the first to succeed in solving cases of the cubic equation that had eluded both the Greek and Islamic mathematicians.

Évariste Galois (1811–1832). The Frenchman Galois possessed one of the most original mathematical minds of all time. Despite his untimely death in a duel, Galois’s work on the solvability of equations eventually overshadowed that of his illustrious contemporaries Abel and Cauchy. He completely resolved the issue of solvability of equations and in the process created the modern mathematical discipline of group theory.

Carl Fridrich Gauss (1777–1855). The German Gauss is generally considered to be the greatest of the mathematicians to follow Newton. His profound work influenced the subsequent evolution of all the major areas of mathematics. In addition, he also made important contributions to astronomy, physics, and geodesy.

Camille Jordan (1838–1922). A French mathematician who is best known for his pioneering work in group theory and linear algebra.

Omar Khayyam (1048–1131). Better known for his collection of poems *Rubaiyat*, the Persian Khayyam also wrote several mathematical, astronomical, and philosophical tracts. He tried to systematize the solution of the cubic equation and actually solved several special cases. He is also known for his work on Euclid's Parallel Postulate.

Felix Klein (1849–1925). Klein was a highly influential German mathematician. His paper that came to be known as the *Erlanger Programm* focused the attention of mathematicians on the applications of group theory to geometry. He was also one of the pioneers of hyperbolic geometry and the calculus of complex variables.

Ernst E. Kummer (1810–1893). Kummer lived and worked in Germany all his life. He began his mathematical career as a high school teacher. His main mathematical work was in the area of algebra where he picked up where Gauss left off on the topic now known as algebraic number theory. He invented the term “ideal element,” which eventually led to the modern formulation in the language of rings and ideals.

Joseph Louis Lagrange (1736–1813). Lagrange was born in Italy but spent most of his adult life in France, from where his ancestors had come. He made vast contributions to the fast evolving disciplines of calculus and differential equations. His work on algebraic equation, linear algebra, and number theory also proved very fecund in the long run.

Adrien-Marie Legendre (1752–1833). Legendre was a highly influential French mathematician. He made substantial contributions to geometry, analysis, and number theory, in each of which areas he also wrote a definitive text.

Isaac Newton (1642–1727). An English scientist whose creativity influenced the evolution of both mathematics and physics more than that of any other individual. Among his many achievements are his theories of light and gravitation, and the development of calculus. His best known book is the *Principia Mathematica*.

Blaise Pascal (1623–1662). Pascal was a French mathematician, philosopher, and scientist. He invented the first adding machine and made major contributions toward the evolution of geometry and the theory of probability.

Joseph Raphson (1648–1715). An Englishman whose tract *Analysis Aequationum Universalis* described Newton's numerical method for solving equations.

Paolo Ruffini (1765–1822). Ruffini was an Italian physician who published a proof of the unsolvability of the general quintic equation by radicals. While his proof was incomplete, it did contain some ideas that were eventually incorporated into Abel's proof of the same theorem.

Niccolò Tartalia (1499–1557). An Italian mathematician whose work on some cubic equations, together with that of his compatriots del Ferro and Cardano, resulted in the complete solution of the cubic equation.

A. T. Vandermonde (1735–1796). A French mathematician whose pioneering work on the algebraic solution of equations was eclipsed by that of Lagrange.

Bibliography

- Abel, N. H., *Oeuvres complètes de Niels Henrik Abel*, Christiana, Gropdahl, 1881.
- Bell, E. T., *Men of Mathematics*, Simon and Schuster, New York, 1965.
- Birkeland, B., *Ludwig Sylow's Lectures on Algebraic Equations and Substitutions*, Christiana, Oslo, 1862: An Introduction and a Summary, *Historia Mathematica*, 23 (1996), 182–199.
- Birkhoff, G., and Mac Lane, S., *A Survey of Modern Algebra*, 4th ed., Macmillan, New York, 1977.
- Bombelli, R., *Algebra*, Feltrinelli, Milan, 1966.
- Borofsky, S., *Elementary Theory of Equations*, Macmillan, New York, 1959.
- Cajori, F., *An Introduction to the Theory of Equations*, Macmillan, New York, 1969.
- Cardano, G, *Ars Magna*, Dover, New York, 1993.
- Cauchy, A. L., *Oeuvres complètes*, Gauthier-Villars, Paris, 1882.
- Cauchy, A. L., Mémoire sur les premiers termes de la série des quantités qui sont propres a représenter le nombre des valeurs distinctes d'une fonction des n variables indépendantes, *Comptes Rendus Paris*, 21 (1845), 1093–1101.
- Cauchy, A. L., Mémoire sur une nouvelle théorie des imaginaires, et sur les racines symboliques des équations et des equivalences, *Comptes Rendus Paris*, 24 (1847), 1120–1130.
- Dickson, L. E., *New First Course in the Theory of Equations*, Wiley, New York, 1939.
- Dickson, L. E., *Linear Groups*, Dover, New York, 1958.
- Edwards, H. M., *Galois Theory*, Springer, New York, 1984.
- Euclid, *The Elements*, Dover, New York, 1956.
- Fermat, P. de, *Oeuvres de Fermat*, ed. P. Tannery and C. Henry, Gauthier-Villars, Paris, 1922.
- Gallian, J. A., *Contemporary Abstract Algebra*, 2nd ed., D. C. Heath, Lexington, MA, 1982.
- Galois, É., *Écrits et mémoires mathématiques*, ed. R. Bourgne and J.-P. Azra, Gauthier-Villars, Paris, 1962.
- Galois, É., Sur la théorie des nombres, *Bulletin des Sciences mathématiques*, 428 (1830).
- Gauss, C. F., translated by Arthur A. Clarke: *Disquisitiones Arithmeticae*, Yale University Press, 1965.

- Gillings, R. J., *Mathematics in the Time of the Pharaohs*, Dover, New York, 1972.
- Hadlock, C. R., *Field Theory and Its Classical Problems*, Mathematical Association of America, Washington, DC, 1978.
- Hahn, L., *Complex Numbers and Geometry*, Mathematical Association of America, Washington, DC, 1994.
- Hall, H. S., and Knight, S. R., *Higher Algebra*, Macmillan, London, 1919.
- Herstein, I. N., *Topics in Algebra*, 2nd ed., Wiley, New York, 1975.
- Hungerford, T. W., *Algebra*, Springer, New York, 1974.
- Jordan, C., *Traité des substitutions et des equations algebriques*, Gauthier-Villars, Paris, 1870.
- Katz, V., *A History of Mathematics: An Introduction*, HarperCollins, New York, 1993.
- Kiernan, B. M., The development of Galois theory from Lagrange to Artin, *Arch. Hist. Exact Sci.*, 8 (1971), 40–154.
- Kleiner, I., The evolution of group theory: a brief survey, *Math. Mag.*, 59 (1986), 195–215.
- Kline, M., *Mathematical Thought from Ancient to Modern Times*, Oxford University Press, New York, 1990.
- Knopp, K., *Problem Book in the Theory of Functions*, Dover, New York, 1948.
- Lagrange, J. L., *Oeuvres de Lagrange*, Gauthier-Villars, Paris, 1867–1892.
- MacWilliams, F. J., and Sloane, N. J. A., *The Theory of Error-Correcting Codes*, North-Holland, Amsterdam, 1977.
- McCarthy, P. J., *Algebraic Extensions of Fields*, Dover, New York, 1991.
- Needham, J., *Science and Civilization in China*, Cambridge University Press, Cambridge, 1959.
- Serret, J. A., *Cours d'algebra superieure*, Gauthier-Villars, Paris, 1885.
- Shanks, D., *Solved and Unsolved Problems in Number Theory*, Spartan Books, Washington, 1962.
- Stauduhar, R. P., The Determination of Galois Groups, *Mathematics of Computation*, 27 (1973), 981–996.
- Stewart, I., *Galois Theory*, Chapman and Hall, London, 1989.
- Story, W. E., Note on the “15” Puzzle, *Amer. J. Math.*, 2 (1879), 399–404.
- van der Waerden, B. L., *Modern Algebra*, F. Ungar, New York, 1953.
- Wells, D., *The Penguin Dictionary of Curious and Interesting Numbers*, Penguin, Great Britain, 1986.
- Wussing, H., *The Genesis of the Abstract Group Concept*, MIT Press, Cambridge, 1984.

Solutions to Selected Exercises

1.1.3 $(1 \pm \sqrt{7})/6$ 1.1.9 $(3abc - b^3)/a^3$ 1.1.11 $(b^2 - 4ac)/a^2$ 1.1.13 $-ab/c^2$
 1.1.15 $x^2 + (p - q)x - pq = 0$ 1.1.17 $\alpha \leq 0$ or $\alpha \geq 4$ 1.1.1 True 1.1.3 False



2.1.1 Argument: $\arctan(3/2) \approx 56.3^\circ$; modulus: $\sqrt{2^2 + 3^2} = \sqrt{13} \approx 3.6$. 2.1.3 Argument:
 $180^\circ + \arctan(4/3) \approx 233.1^\circ$; modulus: $\sqrt{(-3)^2 + (-4)^2} = 5$. 2.1.5 $7 + 2i$ 2.1.7 $-3 + 4i$
 2.1.9 $13 + 13i$ 2.1.11 $\frac{7}{26} + \frac{17}{26}i$ 2.1.13 $-\frac{3\sqrt{3}}{7} + \frac{13}{7}i$ 2.1.15 $-\frac{1}{2} - \frac{7}{2}i$ 2.1.17 $-7 + 24i$
 2.1.19 i 2.1.21 $-i$ 2.1.23 $z = -7 - 3i$ 2.1.25 $w = z = i$
 2.2.1 $\{\frac{1}{\sqrt{2}} + \frac{i}{\sqrt{2}}, i, -\frac{1}{\sqrt{2}} + \frac{i}{\sqrt{2}}, -1, -\frac{1}{\sqrt{2}} - \frac{i}{\sqrt{2}}, -i, \frac{1}{\sqrt{2}} - \frac{i}{\sqrt{2}}, 1\}$
 2.2.3 $\{\frac{\sqrt{3}}{2} + \frac{i}{2}, i, -\frac{\sqrt{3}}{2} + \frac{i}{2}, -\frac{\sqrt{3}}{2} - \frac{i}{2}, -i, \frac{\sqrt{3}}{2} - \frac{i}{2}\}$ 2.2.5 $\pm(2 - i)$
 2.2.7 $\{1.08 + .29i, -.79 + .79i, -.29 - 1.08i\}$ 2.2.9 $\{i, -\frac{\sqrt{3}}{2} - \frac{i}{2}, \frac{\sqrt{3}}{2} - \frac{i}{2}\}$ 2.2.11 $\pm(c + i)$
 2.2.15 $\{-i, (-2 + i)/5\}$ 2.2.17 $\{-i, -3 - 2i\}$ 2.2.23 -1 2.2.25 ζ^{357} 2.3.1 rational
 2.3.3 degree 2 algebraic 2.3.5 rational 2.3.7 not algebraic 2.3.9 degree 2 algebraic
 2.3.11 not algebraic 2.4.7 102, 120, 128, 136, 160, 170, 192 2.4.13 constructible
 2.4.15 constructible 2.4.17 constructible 2.4.19 not known to be constructible 2.5.1 elements
 $i, -1, -i$, and 1 with orders $4, 2, 4$, and 1 , respectively. 2.5.3 elements $-\omega^2, \omega, -1, \omega^2, -\omega$,
 and 1 with orders $6, 3, 2, 3, 6$, and 1 , respectively. 2.5.5 elements $(1 + i)/\sqrt{2}, i, (-1 + i)/\sqrt{2},$
 $-1, (-1 - i)/\sqrt{2}, -i, (1 - i)/\sqrt{2}$, and 1 with orders $8, 4, 8, 2, 8, 4, 8$, and 1 , respectively.
 2.5.7 Let $\zeta = \cos(2\pi/10) + i\sin(2\pi/10)$. Then $\sqrt[10]{1}$ has elements $\zeta, \zeta^2, \zeta^3, \zeta^4, \zeta^5, \zeta^6, \zeta^7, \zeta^8,$
 ζ^9 , and 1 , with orders $10, 5, 10, 5, 2, 5, 10, 5, 10$, and 1 , respectively. 2.5.9 $-\omega^2$ and ω
 2.5.11 ζ, ζ^5, ζ^7 , and ζ^{11} where $\zeta = \cos(2\pi/12) + i\sin(2\pi/12)$ 2.1.1 true 2.1.3 false 2.1.5 true
 2.1.7 false 2.1.9 true 2.1.11 false 2.1.13 true



3.1.1 $x_1 = 3^{2/3} - 9/(3 \cdot 3^{2/3}) = 3^{2/3} - 3^{1/3}, x_2 = \omega 3^{2/3} - \omega^2 3^{1/3}, x_3 = \omega^2 3^{2/3} - \omega 3^{1/3}$
 3.1.3 $x_1 = 6^{2/3} - 18/(3 \cdot 6^{2/3}) = 6^{2/3} - 6^{1/3}, x_2 = \omega 6^{2/3} - \omega^2 6^{1/3}, x_3 = \omega^2 6^{2/3} - \omega 6^{1/3}$
 3.1.5 $x_1 = y_1 - a/3 = 2^{2/3} - 6/(3 \cdot 2^{2/3}) - 1 = 2^{2/3} - 2^{1/3} - 1, x_2 = \omega 2^{2/3} - \omega 2^{1/3} - 1,$
 $x_3 = \omega^2 2^{2/3} - \omega^2 2^{1/3} - 1$
 3.1.7 $x_1 = y_1 - a/3 = -2^{1/3}i\omega - (-6)/(3 \cdot (-2^{1/3}i\omega)) - 1 = -2^{1/3}i\omega + 2^{2/3}i\omega^2 - 1,$
 $x_2 = -2^{1/3}i\omega^2 + 2^{2/3}i\omega - 1, x_3 = -2^{1/3}i + 2^{1/3}i - 1$
 3.1.9 $x_1 = y_1 - 1/2 = 1/2 - (3/4)/(3 \cdot 1/2) - 1/2 = -1/2,$

$x_2 = \omega(1/2) - \omega^2(3/4)/(3 \cdot 1/2) - 1/2 = (\omega - \omega^2 - 1)/2 = \omega$,
 $x_3 = \omega^2(1/2) - \omega(3/4)/(3 \cdot 1/2) - 1/2 = (\omega^2 - \omega - 1)/2 = \omega^2$ 3.1.11 $-1 - i$ 3.3.1 $x_1 = 0$,
 $x_2 = -.6667$, $x_3 = -.5968$, $x_4 = x_5 = -.5958$ 3.3.3 $x_1 = 0$, $x_2 = -.85$, $x_3 = x_4 = -.8449$
3.3.5 $x_1 = 0$, $x_2 = .5667$, $x_3 = x_4 = .5658$ 3.3.7 $x_1 = -1$, $x_2 = -2.8305$, $x_3 = -2.0496$,
 $x_4 = -1.9387$, $x_5 = x_6 = -1.9346$ 3.3.9 $x_1 = 2$, $x_2 = 2.1667$, $x_3 = 2.1545$, $x_4 = x_5 = 2.1544$
3.3.11 $x_1 = 3$, $x_2 = 4.5311$, $x_3 = 4.0488$, $x_4 = 3.7947$, $x_5 = 3.7311$, $x_6 = x_7 = 3.7276$
3.1.1 False 3.1.3 False 3.1.5 True



4.1.1 $x \equiv 1$ 4.1.3 $x \equiv 0, 1$ 4.1.5 $x \equiv 0$ 4.1.7 $x \equiv 4$; $x \equiv 2, 3$; $x \equiv 0, 2, 3$; $x \equiv 4$; $x \equiv 0$
4.1.9 No solutions 4.1.11 $y \equiv 3$, $x \equiv 6$ 4.1.13 $y \equiv 3$, $z \equiv 2$, $x \equiv 3$ 4.1.15 $x \equiv 1$
4.1.17 $x \equiv 0$, $y \equiv 3$; $y \equiv 3$, $x \equiv 11$ 4.1.19 $x \equiv 7$; no solution in \mathbb{Z}_{13} 4.1.21 501 4.1.23 4
4.2.1 365 4.2.3 12 4.2.5 1 4.2.7 59 4.2.9 18 4.2.11 34 4.2.13 72 4.2.15 31
4.2.17 65,521 4.2.19 $1 \equiv 1^{-1}$, $11 \equiv 11^{-1}$, $5 \equiv 5^{-1}$, $7 \equiv 7^{-1}$ 4.2.21 $x = -42$, $y = 24$
4.3.1 Squares: 0, 1, 2, 4 4.3.3 Squares: 0, 1, 3, 4, 9, 10, 12 4.3.5 0, 1, 6 4.3.7 0, 1, 5, 8, 12
4.4.1 2^{65^6} 4.4.3 $641 \cdot 6,700,417$ 4.4.5 2^{20} 4.4.15 325 4.1.1 True 4.1.3 True 4.1.5 False
4.1.7 True



5.1.1 $128x^7 + 1,344x^6y^2 + 6,048x^5y^4 + 15,120x^4y^6 + 22,680x^3y^8 + 20,412x^2y^{10} +$
 $10,206xy^{12} + 2,187y^{14}$ 5.1.3 $3x^{10} + 4y^5z^{15}$ 5.1.5 z^{12} 5.1.7 $7,015,680a^{26}b^8c^{12}$
5.1.9 489,888 5.2.1 0, 1, 2, 4, 4, 2, 1 5.2.3 $x = 1$ 5.2.5 $x = 1$ 5.2.7 $x = 7$ 5.2.15 The
primitive roots mod 11 are the powers 2^n where n is relatively prime to 10, that is, 2, 8, 7, and
6. 5.2.17 For $p = 2$, the root is 1; for $p = 3$, the root is 2; for $p = 5$, the roots are 2 and 4; for
 $p = 7$, the roots are 3 and 5; for $p = 11$, the roots are 2, 6, 7, and 8; for $p = 13$, the roots are 2,
6, 7, and 11; for $p = 17$, the roots are 3, 5, 6, 7, 10, 11, 12, and 14; for $p = 19$, the roots are
2, 3, 10, 13, 14, and 15. 5.2.27 $a = 2$ and $n = 4$ 5.3.1 $x^2 + y^2 + z^2 + 2xy + 2xz + 2yz$
5.3.3 $x^4 + y^4 + z^4 + 4x^3y + 4xy^3 + 4x^3z + 4xz^3 + 4y^3z + 4yz^3 + 6x^2y^2 + 6x^2z^2 + 6y^2z^2 +$
 $12x^2yz + 12xy^2z + 12xyz^2$
5.3.5 $x^6 + y^9 + x^3y^3 + 3x^4y^3 + 3x^2y^6 + 3x^5y + 3x^4y^2 + 3xy^7 + 3x^2y^5 + 6x^3y^4$
5.3.7 887,400 5.3.9 2,355 5.4.1 $\varphi(24) = 8$; $\varphi(144) = 48$; and $\varphi(1,000) = 400$ 5.1.1 True
5.1.3 False 5.1.5 True 5.1.7 False 5.1.9 True



6.1.1 Quotient: $x^4 + x^3 + x^2 + x$; remainder: $x^2 + 1$ 6.1.3 Quotient: $x^4 + 4x^3 + x^2 + 4x + 2$;
remainder: $2x^2 + 2x + 4$ 6.1.11 $x^6 + 2x^3 + 1$ 6.1.13 $x^6 + 2x^5 + 4x^4 + x^3 + 2x^2 + 4x + 1$
6.2.1 1, x , $x + 1$, x^2 , $x^2 + 1 = (x + 1)^2$, $x^2 + x = x(x + 1)$, $x^2 + x + 1$, x^3 ,
 $x^3 + 1 = (x + 1)(x^2 + x + 1)$, $x^3 + x = x(x + 1)^2$, $x^3 + x + 1$, $x^3 + x^2 + 1$, $x^3 + x^2 = x^2(x + 1)$,

$$\begin{aligned}
 x^3 + x^2 + x + 1 &= (x+1)^3, x^4, x^4 + 1 = (x+1)^4, x^4 + x = x(x+1)(x^2 + x + 1), x^4 + x + 1, \\
 x^4 + x^2 &= x^2(x+1)^2, x^4 + x^2 + 1 = (x^2 + x + 1)^2, x^4 + x^2 + x = x(x^3 + x^2 + 1), \\
 x^4 + x^2 + x + 1 &= (x+1)(x^3 + x^2 + 1), x^4 + x^3 = x^3(x+1), x^4 + x^3 + 1, \\
 x^4 + x^3 + x &= x(x^3 + x^2 + 1), x^4 + x^3 + x + 1 = (x+1)^2(x^2 + x + 1), \\
 x^4 + x^3 + x^2 &= x^2(x^2 + x + 1), x^4 + x^3 + x^2 + 1 = (x+1)(x^3 + x + 1), \\
 x^4 + x^3 + x^2 + x &= x(x+1)^3, x^4 + x^3 + x^2 + x + 1 \quad 6.2.3 \quad x^2, x^2 + 1, x^2 + 2 = (x+1)(x+2), \\
 x^2 + x &= x(x+1), x^2 + x + 1 = (x+2)^2, x^2 + x + 2, x^2 + 2x = x(x+2), \\
 x^2 + 2x + 1 &= (x+1)^2, x^2 + 2x + 2 \quad 6.2.5 \quad x^4 + 1 = (x+1)(x^2 + 4x + 1), x^3 + x + 1, \\
 x^3 + 2x + 1, x^3 + 3x + 1 &= (x+4)(x+3)^2, x^3 + 4x + 1 = (x+2)(x^2 + 3x + 3) \\
 6.2.7 \quad p(p^2 - 1)/3 \quad 6.2.13 \quad 3x - 1 \quad 6.2.15 \quad 3x + 1 \quad 6.2.17 \quad a \equiv 8, b \equiv 0 \quad 6.3.1 \quad x^2 + x + 1 \\
 6.3.3 \quad 1 \quad 6.3.5 \quad 1 \quad 6.3.9 \quad \text{No} \quad 6.3.11 \quad \text{No} \\
 6.3.27 \quad (x^3 + 1) &= x^2(x^7 + x^4 + x^3 + 1) + (x^3 + x + 1)(x^6 + x^4 + x + 1) \quad 6.4.1 \quad a^2 - 2b \\
 6.4.3 \quad c - ab \quad 6.4.5 \quad -b/c \quad 6.4.7 \quad (a^2 + b)/(c - ab) \quad 6.4.9 \quad x^3 + 23x - 1 \\
 6.4.11 \quad x^3 - 3x^2 + 26x - 23 \quad 6.4.13 \quad x^3 + 23kx^2 + k^3 \quad 6.4.15 \quad 5 \quad 6.4.17 \quad \pm\sqrt{5}, \pm\sqrt{7}i, 4 \\
 6.4.19 \quad (5 \pm \sqrt{13})/2 \quad 6.4.27 \quad -a_{n-1}/a_n \quad 6.4.29 \quad a_{n-2}/a_n \quad 6.5.1 \quad \pm i, \pm 1 \quad 6.5.2 \quad \omega, \omega^2, 0, 1 \\
 6.1.1 \quad \text{False} \quad 6.1.3 \quad \text{True} \quad 6.1.5 \quad \text{False} \quad 6.1.7 \quad \text{True} \quad 6.1.9 \quad \text{True}
 \end{aligned}$$



$$\begin{aligned}
 7.1.1 \quad \tau, \tau^2, \tau^3 &= \tau + 1, \tau^4 = \tau(\tau + 1) = \tau^2 + \tau, \tau^5 = \tau(\tau^2 + \tau) = \tau^3 + \tau^2 = \tau^2 + \tau + 1, \\
 \tau^6 &= \tau(\tau^2 + \tau + 1) = \tau^3 + \tau^2 + \tau = \tau + 1 + \tau^2 + \tau = \tau^2 + 1, \tau^7 = \tau(\tau^2 + 1) = \tau^3 + \tau = \tau + 1 + \tau = 1, \\
 7.1.3 \quad \eta, \eta^2, \eta^3, \eta^4, \eta^5 &= \eta^2 + 1, \eta^6 = \eta^3 + \eta, \eta^7 = \eta^4 + \eta^2, \eta^8 = \eta^5 + \eta^3 = \eta^3 + \eta^2 + 1, \\
 \eta^9 &= \eta^4 + \eta^3 + \eta, \eta^{10} = \eta^5 + \eta^4 + \eta^2 = \eta^4 + 1, \eta^{11} = \eta^5 + \eta = \eta^2 + \eta + 1, \eta^{12} = \eta^3 + \eta^2 + \eta, \\
 \eta^{13} &= \eta^4 + \eta^3 + \eta^2, \eta^{14} = \eta^5 + \eta^4 + \eta^3 = \eta^4 + \eta^3 + \eta^2 + 1, \\
 \eta^{15} &= \eta^5 + \eta^4 + \eta^3 + \eta = \eta^4 + \eta^3 + \eta^2 + \eta + 1, \eta^{16} = \eta^5 + \eta^4 + \eta^3 + \eta^2 + \eta^1 = \eta^4 + \eta^3 + \eta + 1, \\
 \eta^{17} &= \eta^5 + \eta^4 + \eta^2 + \eta = \eta^4 + \eta + 1, \eta^{18} = \eta^5 + \eta^2 + \eta = \eta + 1, \eta^{19} = \eta^2 + \eta, \eta^{20} = \eta^3 + \eta^2, \\
 \eta^{21} &= \eta^4 + \eta^3, \eta^{22} = \eta^4 + \eta^2 + 1, \eta^{23} = \eta^3 + \eta^2 + \eta + 1, \eta^{24} = \eta^4 + \eta^3 + \eta^2 + \eta, \eta^{25} = \eta^4 + \eta^3 + \eta, \\
 \eta^{26} &= \eta^4 + \eta^2 + \eta + 1, \eta^{27} = \eta^3 + \eta + 1, \eta^{28} = \eta^4 + \eta^2 + \eta, \eta^{29} = \eta^3 + 1, \eta^{30} = \eta^4 + \eta, \\
 \eta^{31} &= \eta^5 + \eta = 1. \\
 7.1.5 \quad \xi, \xi^2 &= \xi + 1, \xi^3 = 2\xi + 1, \xi^4 = 2, \xi^5 = 2\xi, \xi^6 = 2\xi + 2, \xi^7 = \xi + 2, \xi^8 = 1. \quad 7.1.7 \quad \theta, \theta^2, \\
 \theta^3 &= \theta^2 + 2, \theta^4 = \theta^3 + 2\theta = \theta^2 + 2\theta + 2, \theta^5 = 2\theta + 2, \theta^6 = 2\theta^2 + 2\theta, \theta^7 = \theta^2 + 1, \theta^8 = \theta^2 + \theta + 2, \\
 \theta^9 &= 2\theta^2 + 2\theta + 2, \theta^{10} = \theta^2 + 2\theta + 1, \theta^{11} = \theta + 2, \theta^{12} = \theta^2 + 2\theta, \theta^{13} = 2, \theta^{14} = 2\theta, \theta^{15} = 2\theta^2, \\
 \theta^{16} &= 2\theta^2 + 1, \theta^{17} = 2\theta^2 + 2\theta + 1, \theta^{18} = \theta + 1, \theta^{19} = \theta^2 + \theta, \theta^{20} = 2\theta^2 + 2, \theta^{21} = 2\theta^2 + 2\theta + 1, \\
 \theta^{22} &= 2\theta^2 + \theta + 2, \theta^{23} = 2\theta + 1, \theta^{24} = 2\theta^2 + \theta, \theta^{25} = 1. \\
 7.1.9 \quad \alpha^2 &= \alpha + 4, \alpha^3 = 5\alpha + 3, \alpha^4 = 2\alpha + 6, \alpha^5 = \alpha + 1, \alpha^6 = 2\alpha + 4, \alpha^7 = 6\alpha + 1, \alpha^8 = 3, \\
 \alpha^9 &= 3\alpha, \alpha^{10} = 3\alpha + 5, \alpha^{11} = \alpha + 5, \alpha^{12} = 6\alpha + 4, \alpha^{13} = 3\alpha + 3, \alpha^{14} = 6\alpha + 5, \alpha^{15} = 4\alpha + 3, \\
 \alpha^{16} &= 2, \alpha^{17} = 2\alpha, \alpha^{18} = 2\alpha + 1, \alpha^{19} = 3\alpha + 1, \alpha^{20} = 4\alpha + 5, \alpha^{21} = 2\alpha + 2, \alpha^{22} = 4\alpha + 1, \\
 \alpha^{23} &= 5\alpha + 2, \alpha^{24} = 6, \alpha^{25} = 6\alpha, \alpha^{26} = 6\alpha + 3, \alpha^{27} = 2\alpha + 3, \alpha^{28} = 5\alpha + 1, \alpha^{29} = 6\alpha + 6, \\
 \alpha^{30} &= 5\alpha + 3, \alpha^{31} = \alpha + 6, \alpha^{32} = 4, \alpha^{33} = 4\alpha, \alpha^{34} = 4\alpha + 2, \alpha^{35} = 6\alpha + 2, \alpha^{36} = \alpha + 3,
 \end{aligned}$$

$\alpha^{37} = 4\alpha + 4$, $\alpha^{38} = \alpha + 2$, $\alpha^{39} = 3\alpha + 4$, $\alpha^{40} = 5$, $\alpha^{41} = 5\alpha$, $\alpha^{42} = 5\alpha + 6$, $\alpha^{43} = 4\alpha + 6$,
 $\alpha^{44} = 3\alpha + 2$, $\alpha^{45} = 5\alpha + 5$, $\alpha^{46} = 3\alpha + 6$, $\alpha^{47} = 2\alpha + 5$, $\alpha^{48} = 1$. 7.1.11 $y = 1$, $x = 1 + \beta$
 7.1.13 $x = \beta^2$, $y = 1 + \beta + \beta^2$, $z = 1$ 7.1.15 $y = 2$, $x = \beta - 2$ 7.1.17 $y = \beta + 2$, $z = \beta + 2$,
 $x = 0$ 7.1.19 See Theorem 6.2 7.2.1 Each element except 0 and 1 has order 7. 7.2.3 Each
 element except 0 and 1 has order 31. 7.2.5 μ , μ^5 , μ^7 , μ^{11} , μ^{13} , μ^{17} , μ^{19} , and μ^{23} have order 24;
 μ^2 , μ^{10} , μ^{14} , and μ^{22} have order 12; μ^3 , μ^9 , μ^{15} , and μ^{21} have order 8; μ^4 and μ^{20} have order 6;
 μ^6 and μ^{18} have order 4; μ^8 and μ^{16} have order 3; μ^{12} has order 2; 1 has order 1. 7.2.13 0
 except that when $p = 2$ and $v = 1$ the answer is 1. 7.3.1 α and $\alpha^2 = 1 + \alpha$ 7.3.3 2 and 3
 7.3.5 σ , σ^2 , $\sigma^4 = \sigma + 1$, $\sigma^7 = \sigma^3 + \sigma + 1$, $\sigma^8 = \sigma^2 + 1$, $\sigma^{11} = \sigma^3 + \sigma^2 + \sigma$, $\sigma^{13} = \sigma^3 + \sigma^2 + 1$, and
 $\sigma^{14} = \sigma^3 + 1$ 7.3.7 σ , $\sigma^3 = 2\sigma + 2$, $\sigma^5 = 2\sigma$, and $\sigma^7 = \sigma + 1$ 7.3.9 α , $\alpha^5 = \alpha + 1$, $\alpha^7 = 6\alpha + 1$,
 $\alpha^{11} = \alpha + 5$, $\alpha^{13} = 3\alpha + 3$, $\alpha^{17} = 2\alpha$, $\alpha^{19} = 3\alpha + 1$, $\alpha^{23} = 5\alpha + 2$, $\alpha^{25} = 6\alpha$, $\alpha^{29} = 6\alpha + 6$,
 $\alpha^{31} = \alpha + 6$, $\alpha^{35} = 6\alpha + 2$, $\alpha^{37} = 4\alpha + 4$, $\alpha^{41} = 5\alpha$, $\alpha^{43} = 4\alpha + 6$, $\alpha^{47} = 2\alpha + 5$
 7.4.1 $x^2 + 4x + 2$, $x^2 + 3x + 3$, $x^2 + x + 2$, $x^2 + 2x + 3$ 7.4.3 $x^3 + 2x + 1$, $x^3 + 2x^2 + x + 1$,
 $x^3 + x^2 + 2x + 1$, $x^3 + 2x^2 + 1$ 7.4.11 $x^4 + x + 1$ is the minimal polynomial for σ , σ^2 , σ^4 , and
 σ^8 ; $x^4 + x^3 + x^2 + x + 1$ is the minimal polynomial for σ^3 , σ^6 , σ^9 , and σ^{12} ; $x^2 + x + 1$ is the
 minimal polynomial of σ^5 and σ^{10} ; $x^4 + x^3 + 1$ is the minimal polynomial for σ^7 , σ^{14} , σ^{13} , and
 σ^{11} ; x and $x + 1$ are the minimal polynomials of 0 and 1, respectively. 7.4.13 Using
 Exercise 7.1.8, $x^2 + 4x + 2$ is the minimal polynomial for μ and μ^5 ;
 $(x - \mu^2)(x - \mu^{10}) = x^2 + 3x + 4$; $(x - \mu^3)(x - \mu^{15}) = x^2 + 3$; $(x - \mu^4)(x - \mu^{20}) = x^2 + 4x + 1$;
 $(x - \mu^7)(x - \mu^{11}) = x^2 + 3x + 3$; $(x - \mu^8)(x - \mu^{16}) = x^2 + x + 1$; $(x - \mu^9)(x - \mu^{21}) = x^2 + 2$;
 $(x - \mu^{13})(x - \mu^{17}) = x^2 + x + 2$; $(x - \mu^{14})(x - \mu^{22}) = x^2 + 2x + 4$;
 $(x - \mu^{19})(x - \mu^{23}) = x^2 + 2x + 3$; $x - a$ is the minimal polynomial for all $a \in \mathbb{Z}_5$.
 7.1.1 True 7.1.3 False 7.1.5 False 7.1.7 True



8.1.1 One 8.1.3 One 8.1.5 One 8.1.7 One 8.1.9 Two 8.1.11 Three 8.1.13 One
 8.1.15 Three 8.1.17 Six 8.1.19 Four 8.1.21 Six 8.1.23 Six 8.1.25 24 8.1.27 One
 8.1.29 One 8.1.31 20 8.1.33 15 8.1.35 120 8.1.37 $x_1 x_2 \cdots x_n$ 8.1.39 $x_1^2 x_2^2 x_3 \cdots x_n$
 8.1.41 $(x_1 x_2 + x_3 x_4) x_5 x_6 \cdots x_n$ 8.1.43 $x_1 x_2 \cdots x_{n-1} x_n^2$ 8.2.1 (1 9 8 6 5)(2 3 4 7)
 8.2.3 (1 9)(2 8)(3 7)(4 6)(5) 8.2.5 (1)(2), (1 2) 8.2.7 (1), (1 2), (1 3), (1 4),
 (2 3), (2 4), (3 4), (1 2 3), (1 3 2), (1 2 4), (1 4 2), (1 3 4), (1 4 3), (2 3 4), (2 4 3),
 (1 2)(3 4), (1 3)(2 4), (1 4)(2 3), (1 2 3 4), (1 2 4 3), (1 3 2 4), (1 3 4 2), (1 4 2 3),
 (1 4 3 2) 8.2.9 (1 3 7)(2 4 9 6)(5)(8); order is 12. 8.2.11 (1 7)(8 3)(2 5 4)(6)(9); order
 is 6. 8.2.13 (1 5 2)(3 4 7 9)(6 8); order is 12. 8.2.15 (1 2)(2 3)(4 5)(5 6)(6 7)(8 9)
 8.2.17 (1 9)(9 8)(8 6)(6 5)(2 3)(3 4)(4 7), 8.3.1 Yes, x_1 8.3.3 Yes, $x_1 - x_2$ 8.3.5 Yes,
 $\Delta_3 = (x_1 - x_2)(x_1 - x_3)(x_2 - x_3)$. 8.3.7 Yes, $x_1 x_2 x_3 x_4$ 8.3.9 Yes, $x_1 x_2 + x_3 x_4$ 8.3.11 Yes,
 $x_1 x_2 x_3 x_4 x_5$ 8.3.13 No, by Corollary 8.11 8.3.15 Yes, x_3 8.3.17 Yes, $x_1 x_2 x_3 x_4 x_5 x_6$ 8.3.19 No,
 by Theorem 8.10 with $p = 5$ 8.3.21 No, by Theorem 8.10 with $p = 7$ 8.3.23 Yes, $x_1 x_2$

8.3.25 Yes, $(x_1x_2 + x_3x_4)x_5x_6 \cdots x_n$ 8.4.1 Odd 8.4.3 Even 8.4.5 $(1)(2)$ 8.4.7 $(1), (123), (132), (124), (142), (134), (143), (234), (243), (12)(34), (13)(24), (14)(23)$ 8.4.9 $(12), (13), (23)$ 8.4.23 Yes, $x_n\Delta_{n-1}$ 8.4.25 Yes, $x_{n-2}x_{n-1}x_n\Delta_{n-3}$ if $n \geq 5$; $x_4\Delta_3$ if $n = 4$; Δ_3 if $n = 3$ 8.r.1 False 8.r.3 True 8.r.5 True 8.r.7 False 8.r.9 False



9.1.1 Yes 9.1.3 No 9.1.5 No 9.1.7 Id and (12) 9.1.9 All of S_3 9.1.11 Id and (23) 9.1.13 Id, $(12)(34)$, $(13)(24)$, and $(14)(23)$ 9.1.15 Id, (12) , (34) , $(12)(34)$, $(13)(24)$, $(14)(23)$, (1324) , and (1423) 9.1.19 After Id, the rotations are grouped by the nature of their axes. Axis joining vertices: (5462) , $(56)(24)$, (5264) , (1635) , $(13)(56)$, (1536) , (1234) , $(13)(24)$, (1432) ; axis joining midpoints of edges: $(12)(34)(56)$, $(14)(23)(56)$, $(16)(35)(24)$, $(15)(36)(24)$, $(13)(26)(45)$, $(13)(25)(46)$; axis joining centers of opposite faces: $(126)(534)$, $(162)(543)$, $(164)(235)$, $(146)(253)$, $(145)(263)$, $(154)(236)$, $(125)(463)$, $(152)(436)$. 9.1.21 The icosahedron has 10 pairs of opposite faces. The line joining the center of such a pair of faces acts as the axis of two nontrivial rotations of angles $\pm 120^\circ$. This accounts for 20 symmetries. The icosahedron has 15 pairs of opposite edges. The line joining the midpoints of such an edge acts as the axis of a 180° rotation. This accounts for 15 symmetries. The icosahedron has six pairs of opposite vertices. The line joining each such pair of vertices acts as the axis of four nontrivial rotations of angles 72° , 144° , 216° , and 288° , respectively. This accounts for 24 symmetries. Including the identity, we thus have 60 symmetries. 9.1.23 (134) ; 120° clockwise rotation about altitude from 2 9.1.25 (123) ; 120° clockwise rotation about altitude from 4 9.1.27 $(14)(23)$; 180° rotation about the line joining the centers of edges 14 and 23 9.1.29 $(136)(475)$; 120° rotation about the diagonal joining 2 and 8 9.1.31 $(17)(26)(35)(48)$; 120° rotation about the axis joining the midpoints of the edges 26 and 48 9.1.33 $(14)(28)(35)(67)$; 120° rotation about the axis joining the midpoints of the edges 14 and 67 9.1.35 $2n$ 9.2.1 Yes 9.2.3 No 9.2.5 No 9.2.7 No 9.2.9 No 9.2.11 Yes 9.2.13 Yes 9.2.15 $\{0\}, \{1,3\}, \{2\}$ 9.2.17 $\{1\}, \{i, -i\}, \{-1\}$ 9.2.19 Each element is its own inverse. 9.2.21 $\{0\}, \{1,4\}, \{2,3\}$ 9.2.23 $\{1\}, \{5\}$ 9.2.25 $\{0\}, \{1,5\}, \{2,4\}, \{3\}$ 9.2.27 $\{1\}, \{a, e\}, \{b, f\}, \{c, g\}, \{d\}$ 9.3.3 No, K and $(\mathbb{Z}_4, +)$ are not isomorphic to each other. 9.3.11 No 9.3.21 Define $f(z) = z^2$. 9.4.1 $m = 1: \{0\}$; $m = 2: \{0\}, \mathbb{Z}_2$; $m = 3: \{0\}, \mathbb{Z}_3$; $m = 4: \{0\}, \{0,2\}, \mathbb{Z}_4$; $m = 5: \{0\}, \mathbb{Z}_5$; $m = 6: \{0\}, \{0,3\}, \{0,2,4\}, \mathbb{Z}_6$; $m = 7: \{0\}, \mathbb{Z}_7$; $m = 8: \{0\}, \{0,4\}, \{0,2,4,6\}, \mathbb{Z}_8$; $m = 9: \{0\}, \{0,3,6\}, \mathbb{Z}_9$; $m = 10: \{0\}, \{0,5\}, \{0,2,4,6,8\}, \mathbb{Z}_{10}$ 9.4.3 $\{\text{Id}\}, \{\text{Id}, (12)\}, \{\text{Id}, (123), (132)\}, \{\text{Id}, (1234), (13)(24), (1432)\}, S_3, D_4, A_4$, and S_4 , respectively 9.4.5 $\{\text{Id}, (1234), (12)(345), (12), (345), (354)\}, D_4, D_5$, and A_4 , respectively 9.4.7 $\{0\}, \{0, \beta^k\}$ for $k \in \{0, 1, \dots, 6\}$; $\{0, \beta^i, \beta^j, \beta^i + \beta^j\}$ where i and j are distinct elements of $\{0, 1, \dots, 6\}$; $(\text{GF}(2, x^3 + x^2 + 1), +)$ 9.4.9 $\{1\}, \{1, d\}$,

$\{1, a, d, e\}$, $\{1, b, d, f\}$, $\{1, c, d, g\}$, and the whole group 9.4.11 also $\{1, 4, 7, 10, 13\}$ and $\{2, 5, 8, 11, 14\}$ 9.4.13 also $\{1, 10\}$, $\{2, 11\}$, $\{3, 12\}$, $\{4, 13\}$, $\{5, 14\}$, $\{6, 15\}$, $\{7, 16\}$, and $\{8, 17\}$ 9.4.15 also $\{(1\ 3), (1\ 2\ 3), \{(2\ 3), (1\ 3\ 2)\}\}$
 9.4.17 $\{\text{Id}, (1\ 2)(3\ 4), (1\ 3)(2\ 4), (1\ 4)(2\ 3)\}$, $\{(1\ 2\ 3), (1\ 3\ 4), (2\ 4\ 3), (1\ 4\ 2)\}$, $\{(1\ 3\ 2), (2\ 3\ 4), (1\ 2\ 4), (1\ 4\ 3)\}$ 9.4.19 also $\{a, e\}$, $\{b, f\}$, and $\{c, g\}$ 9.4.21 H and $G - H$ 9.4.23 $k = 1: x_1x_2x_3$; $k = 2: (x_1 - x_2)(x_1 - x_3)(x_2 - x_3)$; $k = 3: x_1$; $k = 6: x_1x_2^2x_3^3$; no such function exists for $k = 4$, $k = 5$, or $k > 6 = 3!$. 9.4.25 (a) $k = 1: x_1x_2x_3x_4x_5$; $k = 2: \Delta_5$; $k = 5: x_1$; no such function for $k = 3$ or $k = 4$. (b) $k = 10: x_1x_2$; $k = 12$: let f be the function such that $D_5 = S_{5,f}$; no such function for $k = 11$, $k = 13$, or $k = 14$. 9.4.29 No 9.4.31 S_4
 9.4.33 A_4 9.4.35 D_4 9.4.37 $\{\text{Id}, (1\ 2\ 3\ 4\ 5), (1\ 3\ 5\ 2\ 4), (1\ 4\ 2\ 5\ 3), (1\ 5\ 4\ 3\ 2)\}$
 9.4.47 $\{1, d\}$ 9.4.49 A_4 9.4.51 S_n : if $n = 1$ or 2 , then the center equals S_n . For $n > 2$, the center of S_n is $\{\text{Id}\}$. 9.5.3 $(\mathbb{Z}_4, +)$ 9.5.5 $(\mathbb{Z}_4, +)$ 9.5.7 K 9.5.9 $(\mathbb{Z}_4, +)$ 9.5.11 K
 9.5.15 n 9.5.17 10 9.5.19 5 9.5.21 21 9.5.23 4 9.6.1 $P_0 = \text{Id}$; $P_1 = (0\ 1)$
 9.6.3 $P_0 = \text{Id}$; $P_1 = (0\ 1\ 2\ 3)$; $P_2 = (0\ 2)(1\ 3)$; $P_3 = (0\ 3\ 2\ 1)$ 9.6.5 $P_0 = \text{Id}$;
 $P_1 = (0\ 1\ 2\ 3\ 4)$; $P_2 = (0\ 2\ 4\ 1\ 3)$; $P_3 = (0\ 3\ 1\ 4\ 2)$; $P_4 = (0\ 4\ 3\ 2\ 1)$ 9.6.7 $P_0 = \text{Id}$;
 $P_1 = (0\ 1\ 2\ 3\ 4\ 5)$; $P_2 = (0\ 2\ 4)(1\ 3\ 5)$; $P_3 = (0\ 3)(1\ 4)(2\ 5)$; $P_4 = (0\ 4\ 2)(1\ 5\ 3)$;
 $P_5 = (0\ 5\ 4\ 3\ 2\ 1)$ 9.1.1 True 9.1.3 False 9.1.5 True 9.1.7 True 9.1.9 True 9.1.11 False
 9.1.13 True 9.1.15 True 9.1.17 True 9.1.19 True



10.1.1 $(\mathbb{Z}_4, +)$ 10.1.3 $(\mathbb{Z}_5, +)$ 10.1.5 H is not normal in G 10.1.7 $(\mathbb{Z}_3, +)$ 10.1.9 H is not normal in G 10.1.11 $(\mathbb{Z}_4, +)$ 10.1.13 $G/\langle (1\ 2\ 3\ 4) \rangle \cong (\mathbb{Z}_2, +)$; $G/\langle \text{Id}, (1\ 3)(2\ 4) \rangle \cong K$.
 10.1.15 $G/\langle x \rangle \cong K$ if $x \neq 0$; $G/\langle 0, x, y, x + y \rangle \cong (\mathbb{Z}_2, +)$ if x and y are distinct and nonzero.
 10.1.17 The nontrivial subgroups of G are $\{1, \sigma^2, \sigma^4, \sigma^6\}$ and $\{1, \sigma^4\}$. The quotients are isomorphic to $(\mathbb{Z}_2, +)$ and $(\mathbb{Z}_4, +)$, respectively. 10.1.19 $G/K \cong (\mathbb{Z}_3, +)$. 10.1.29 It follows from Exercise 8.2.23 that each conjugacy class consists of a maximal set of permutations whose disjoint cycle decompositions all have the same number of k -cycles for each positive integer k .
 10.3.1 $\{a + b\sqrt{5} \mid a, b \in \mathbb{Q}\}$ 10.3.3 $\{a + bi \mid a, b \in \mathbb{Q}\}$
 10.3.5 $\{a + b\omega \mid a, b \in \mathbb{Q}\} = \{a + b\sqrt{-3} \mid a, b \in \mathbb{Q}\}$
 10.3.7 $\{a + b\sqrt{2} + ci + di\sqrt{2} \mid a, b, c, d \in \mathbb{Q}\}$ 10.3.9 $\text{GF}(2, x^2 + x + 1)$
 10.3.11 $\text{GF}(2, x^4 + x^3 + x^2 + x + 1)$ 10.3.13 This is $\text{GF}(2, x^2 + x + 1)$, which has order 4.
 10.3.15 This is $\text{GF}(2, x^4 + x^3 + x^2 + x + 1)$, which has order 16. 10.3.19 $-\frac{2}{41} + \frac{3}{41}[x]$
 10.3.20 $\frac{1}{6} - \frac{1}{6}[x]$ 10.1.1 False 10.1.3 False 10.1.5 True



11.2.9 Groups (a), (d), (e), and (g) are all cyclic of order 8. Group (f) has every element but the identity of order 2, so it is isomorphic to group (i). Group (h) is commutative with an element of order 4 but none of order 8. Groups (b) and (c) are noncommutative and also not isomorphic to

each other. 11.2.11 Groups (a), (c), (d), and (e) are all cyclic and so isomorphic to each other. Groups (i) and (d) are isomorphic to each other. 11.1.1 False 11.1.3 False 11.1.5 True 11.1.7 True



12.1.1 (a) If neither is divisible by 3 then $x^2 \equiv 1 \pmod{3}$. It follows that $z^2 = x^2 + y^2 \equiv 1 + 1 = 2$. This, however, is impossible. (b) Suppose none of x, y, z is divisible by 5. Then x^2, y^2, z^2 are each in the set $\{\pm 1\} \pmod{5}$. Consequently $x^2 + y^2 \in \{\pm 2, 0\}$ which is disjoint from $\{\pm 1\}$. Hence $x^2 + y^2 \neq z^2$. 12.1.3 Primitive: (319, 360, 481), (31, 480, 481), (600, 481, 769). Non primitive: (156, 455, 481), (481, 3108, 3145), (481, 8892, 8905). 12.1.5 Primitive: (5, 12, 13). Non-primitive: (6, 8, 10) 12.2.1 $a^{1/2}$ is interpreted as \sqrt{a} . If $p = 2$ then a is a quadratic residue modulo 2 if and only $a \equiv 1$ iff $\sqrt{a} \equiv 1$, which is obvious. 12.2.3 Note that $5525 = 5 \cdot 5 \cdot 13 \cdot 17$ and each of the primes 5, 13, 17 is the sum of two squares. Several applications of Brahmagupta's proposition then yield $5255 = 14^2 + 73^2$. 12.2.5 Suppose that $p \equiv 1 \pmod{4}$. Then $(-1)^{((p-1)/2)} \equiv 1$, so that $(-1)^{((p-1)/2)}$ is a solution of $x^2 + 1 \equiv 1 \pmod{4}$. 12.2.7 The only prime divisor of n of form $4k + 3$ is 3 and it has an even exponent of 2. Hence, by Proposition 12.2.5, the required triangle exists. 12.3.1 Let p be an odd prime and a not be a multiple of p . Then, modulo p , the following are equivalent for any number x : $ax^2 + bx + c \equiv 0$ and $(2ax + b)^2 - (b^2 - 4ac) \equiv 0$. Consequently the first equation has 0, 1, or 2 distinct solutions according as $b^2 - 4ac$ is a quadratic nonresidue, 0, or a quadratic residue. 12.3.3 1, 4, 9, 16, 25, 5, 18, 20, 14, 10, 8 12.3.5 Pair each quadratic residue a different from 1 and -1 with its multiplicative inverse a^* where $aa^* \equiv 1 \pmod{p}$. Since the product of each pair is equivalent to 1 (mod p) it follows that the product of the quadratic residues of p is congruent to $-1 \pmod{p}$ is -1 if -1 is a quadratic residue of p and to 1 otherwise. 12.3.7 First suppose that $p = 4k + 1$. By the Law of Quadratic Reciprocity $(7/p)(p/7) = 1$. However, this is $(r/7)$ where r is the remainder when p is divided by 7 and it is easy to check that $(r/7) = 1$ for $r = 1, 2, 4$. Thus, p must be of the form $28k + 1$, $28k + 9$ or $28k + 25$. 12.3.9 (a) 1, (b) 1, (c) -1 , (d) -1 12.3.11 Let S denote the given sum. Then $6S = ((p-1)/2)((p+1)/2)p \equiv 0 \pmod{p}$ and hence $p \mid 6S$. Since p is neither 2 nor 3, we have that $p \mid S$. 12.4.1 13, 21, 21, 21, 25 12.4.5 $x = 0, y = 1$ 12.4.7 $(2+i)(3-i)(5-2i)$ 12.4.9 $(2+i)(3+i)^2(7-2i)^2$ 12.4.11 $8-2i = (2-i)(3+i) + 1-i$ 12.4.13 (a) $3+i$ (c) $3+4i$ 12.5.1 11, 11, 15, 15, 17 12.5.7 $(1+\sqrt{-2})(2+5\sqrt{-2})$ 12.5.9 $(1+\sqrt{-2})(2+5\sqrt{-2})(5-\sqrt{-2})$ 12.5.11 $r = 1-2\sqrt{-2}, g = 3+\sqrt{-2}$ 12.5.13 (a) $3+\sqrt{-2}$, (c) $3+4\sqrt{-2}$



13.2.1 0, 1, 2, 3, $\sqrt{-5}$, $1+\sqrt{-5}$, $2+\sqrt{-5}$, $3+\sqrt{-5}$, $4+\sqrt{-5}$, $1+2\sqrt{-5}$ 13.2.3 Irreducibles of norm < 30 : 0, 1, 2, 3, $\sqrt{-5}$, $1+\sqrt{-5}$, $2+\sqrt{-5}$, $3+\sqrt{-5}$, $4+\sqrt{-5}$,

p	$-10 \equiv (\text{mod } p)$	\mathfrak{p}_p	$\langle p \rangle$	$N(\mathfrak{p}_p)$
2	$\equiv 0$	$\langle \sqrt{-10}, 2 \rangle$	\mathfrak{p}_2^2	2
3	$\equiv 2$	$\langle 3 \rangle$	$\langle 3 \rangle$	9
5	$\equiv 0$	$\langle \sqrt{-10}, 5 \rangle$	\mathfrak{p}_5^2	5
7	$\equiv 4$	$\langle 2 + \sqrt{-10}, 7 \rangle$	$\mathfrak{p}_7 \overline{\mathfrak{p}}_7$	7
11	$\equiv 1$	$\langle 1 + \sqrt{-10}, 11 \rangle$	$\mathfrak{p}_{11} \overline{\mathfrak{p}}_{11}$	11
13	$\equiv 3$	$\langle 4 + \sqrt{-10}, 13 \rangle$	$\mathfrak{p}_{13} \overline{\mathfrak{p}}_{13}$	13
17	$\equiv 7$	$\langle 17 \rangle$	$\langle 17 \rangle$	289
19	$\equiv 9$	$\langle 3 + \sqrt{-10}, 19 \rangle$	$\mathfrak{p}_{19} \overline{\mathfrak{p}}_{19}$	19
23	$\equiv 13$	$\langle 6 + \sqrt{-10}, 23 \rangle$	$\mathfrak{p}_{23} \overline{\mathfrak{p}}_{23}$	23

Table S.1 Solution to Exercise 13.7.1

$1 + 2\sqrt{-5}$, $3 + 2\sqrt{-5}$. Others: $4 = 2 \cdot 2$; $5 = \sqrt{-5} \cdot -\sqrt{-5}$; $2\sqrt{-5} = 2 \cdot \sqrt{-5}$;
 $2 + 2\sqrt{-5} = 2 \cdot (1 + \sqrt{-5})$. 13.3.1 (a) True, (b) True, (c) False, (d) True, (e) False, (f) True, (g)
True 13.3.3 (a) Principal: $\langle 9 \rangle$ (b) Principal: $\langle 2 - \sqrt{-6} \rangle$ (c) Principal: $\langle 2 \rangle$ (d) Principal $\langle 3 \rangle$ (e)
Not Principal (f) Principal: $\langle 1 \rangle$ (g) Principal: $\langle 1 \rangle$ (h) Principal: $\langle 1 \rangle$ (i) Principal $\langle 1 \rangle$ (j) Principal:
 $\langle 1 \rangle$ 13.3.5 Suppose $N(a + b\sqrt{-6}) = 2$. $a^2 + 6b^2 = 2$. $b^2 \leq \frac{1}{3} \Rightarrow b = 0$. $a^2 = 2$ has no solution
for a an integer. 13.3.7 $p\overline{p} = \langle 5, 2 + \sqrt{-6} \rangle \cdot \langle 5, 2 - \sqrt{-6} \rangle = \langle 25, 10 - 5\sqrt{-6}, 10 + 5\sqrt{-6}, 10 \rangle$.
5 divides each generator of $p\overline{p}$, so $p\overline{p} \subset \langle 5 \rangle$. $25 - (2) \cdot 10 = 5$, so $5 \in p\overline{p}$ and $\langle 5 \rangle \subset p\overline{p}$.
 $q\overline{q} = \langle 4, 4 - 2\sqrt{-6}, 4 + 2\sqrt{-6}, 10 \rangle$. 2 divides each generator: $p\overline{p} \subset \langle 2 \rangle$. $2 = 10 - (2)4 \in p\overline{p}$, so
 $\langle 2 \rangle \subset p\overline{p}$. $pq = \langle 10, 10 + 5\sqrt{-6}, 4 + 2\sqrt{-6}, (2 + \sqrt{-6})^2 \rangle$. Since $10 = (2 + \sqrt{-6})(2 - \sqrt{-6})$,
each generator is divisible by $2 + \sqrt{-6}$. $(10 + 5\sqrt{-6}) - 2(4 + 2\sqrt{-6}) = 2 + \sqrt{-6} \in pq$.
13.4.1 (a) 36 or -36, (b) 36 or -36, (c) 1 or -1, (d) 1 or -1
13.4.3 Suppose it is principal; that is, there exists α such that $\alpha \cdot \beta = 3$ and $\alpha \cdot \gamma = 1 + \sqrt{-5}$.
Then $N(\alpha) \cdot N(\beta) = 9$ and $N(\alpha) \cdot N(\gamma) = 6$. However, no element of $\mathbb{Z}[\sqrt{-5}]$ has norm 3. Thus
 $N(\alpha) = 1$.
Consider $3(a + b\sqrt{-5}) + (1 + \sqrt{-5})(c + d\sqrt{-5}) = 1$. This gives two equations of rational
integers: $3a + c - 5d = 1$ and $3b + c + d = 1$. Then $c - 5d \equiv 1 \pmod{3}$ and $c + d \equiv 0 \pmod{3}$.
Subtracting the equations gives $-6d \equiv 1 \pmod{3}$. $1 \notin \langle 3, 1 + \sqrt{-5} \rangle$. 13.5.1 (a) 36, (b) 36, (c) 1,
(d) 1 13.6.3 $\langle 5 \rangle$ is not prime. $\langle 5, 2 + \sqrt{-6} \rangle \langle 5, 2 - \sqrt{-6} \rangle = \langle 5 \rangle$. 13.6.5 $\langle 6 \rangle$ is not prime.
 $\langle 2 \rangle \cdot \langle 3 \rangle = \langle 6 \rangle$. 13.7.1 See Table S.1 13.7.3 See Table S.2 13.7.5 Let α and β be rational
primes. Suppose $\alpha, \beta \in I$ and α, β are not associates. Then $\langle \alpha, \beta \rangle = \langle 1 \rangle$ since α, β are relatively
prime rational integers. $\alpha, \beta \in I$ means $I \mid \langle \alpha, \beta \rangle$, so $I = \langle 1 \rangle$.

p	$-14 \equiv (\text{mod } p)$	\mathfrak{p}_p	$\langle p \rangle$	$N(\mathfrak{p}_p)$
2	$\equiv 0$	$\langle \sqrt{-14}, 2 \rangle$	\mathfrak{p}_2^2	2
3	$\equiv 1$	$\langle 1 + \sqrt{-14}, 3 \rangle$	$\mathfrak{p}_3 \overline{\mathfrak{p}}_3$	3
5	$\equiv 1$	$\langle 1 + \sqrt{-14}, 5 \rangle$	$\mathfrak{p}_5 \overline{\mathfrak{p}}_5$	5
7	$\equiv 0$	$\langle \sqrt{-14}, 7 \rangle$	\mathfrak{p}_7^2	7
11	$\equiv 8$	$\langle 11 \rangle$	$\langle 11 \rangle$	121
13	$\equiv 12$	$\langle 5 + \sqrt{-14}, 13 \rangle$	$\mathfrak{p}_{13} \overline{\mathfrak{p}}_{13}$	13
17	$\equiv 3$	$\langle 17 \rangle$	$\langle 17 \rangle$	289
19	$\equiv 5$	$\langle 9 + \sqrt{-14}, 19 \rangle$	$\mathfrak{p}_{19} \overline{\mathfrak{p}}_{19}$	19
23	$\equiv 9$	$\langle 3 + \sqrt{-14}, 23 \rangle$	$\mathfrak{p}_{23} \overline{\mathfrak{p}}_{23}$	23

Table S.2 Solution to Exercise 13.7.3

$$13.7.6 \quad \langle 5, -2 + \sqrt{-1} \rangle \cdot \langle 5, -2 - \sqrt{-1} \rangle = \langle 5 \cdot 5, 5 \cdot (-2 - \sqrt{-1}), 5 \cdot (-2 + \sqrt{-1}), (-2 + \sqrt{-1})(-2 - \sqrt{-1}) \rangle = \langle 5 \cdot 5, 5 \cdot (-2 - \sqrt{-1}), 5 \cdot (-2 + \sqrt{-1}), 5 \rangle = \langle 5 \rangle.$$

28

14.1.1 Let $x = a \cdot b$ be rational integers. x is prime if and only if $x \mid a$ or $x \mid b$. Clearly $a \mid x$, so $x \mid a$ if and only if $\langle x \rangle = \langle a \rangle$ if and only if b is a unit. a or b is a unit if and only if x is irreducible.

14.1.3 Suppose n is divisible by a^2 for some rational integers n and a , nonunits. Consider $\frac{n}{a}$, which is a rational integer since $a \mid n$. $n \nmid \frac{n}{a}$ and $(\frac{n}{a})^2 = \frac{n^2}{a^2} = n \cdot \frac{n}{a^2}$. Since $a^2 \mid n$, this last fraction is still a rational integer, which shows that $n \mid (\frac{n}{a})^2$. Suppose $n \nmid m$ and $n \mid m^2$. Let $m = p_1^{a_1} p_2^{a_2} \dots p_t^{a_t}$ be the prime factorization of m ; $m^2 = p_1^{2a_1} p_2^{2a_2} \dots p_t^{2a_t}$. Since $n \mid m^2$, n has no prime factors which do not appear in the list $\{p_1, \dots, p_t\}$. Since $n \nmid m$, at least one of these factors must appear with power greater than 1. Thus, there is a square which divides n .

14.1.5 By part (a) of Proposition 14.2, we have $(-a) \cdot b = ((-1) \cdot a) \cdot b$. Associativity gives $= (-1) \cdot (a \cdot b) = -(a \cdot b)$ which is the third term. Commutativity gives $= (a \cdot (-1)) \cdot b = a \cdot ((-1) \cdot b) = a \cdot (-b)$. 14.1.7 Suppose b and c are both a multiplicative inverse of a . That is, $a \cdot b = a \cdot c = 1$. $a \cdot b = 1$ if and only if $c \cdot (a \cdot b) = c \cdot 1 = c$ if and only if $(c \cdot a) \cdot b = c$ if and only if $1 \cdot b = c$ if and only if $b = c$.

14.1.9 Ring. Units are 1 and 2 since $2 \cdot 2 = 1$. 14.1.11 Ring. Units are 1 and 5. $2 \cdot 3 = 4 \cdot 3 = 0$; $5 \cdot 5 = 1$. 14.1.13 Not a ring; subtraction is not commutative.

$3 - 1 \neq 1 - 3$. 14.1.15 Not a ring; cross product is not commutative and has no identity.

14.1.17 Not a ring; no additive (union-wise) inverse. 14.1.19 Ring. Units are $f : \mathbb{R} \rightarrow \mathbb{R}$ such that $f(x) \neq 0$ for all $x \in \mathbb{R}$. 14.2.1 (a) $\langle 3, 1 + 2\sqrt{-5} \rangle \langle 3, 1 - 2\sqrt{-5} \rangle = \langle 9, 3 + 6\sqrt{-5}, 3 -$

+	1	-1-i	-i	1-i	2-i	-1+i	i	1+i	-1
1	-1-i	-i	1-i	2-i	-1+i	i	1+i	-1	0
-1-i	-i	1-i	2-i	-1+i	i	1+i	-1	0	1
-i	1-i	2-i	-1+i	i	1+i	-1	0	1	-1-i
1-i	2-i	-1+i	i	1+i	-1	0	1	-1-i	-i
2-i	-1+i	i	1+i	-1	0	1	-1-i	-i	1-i
-1+i	i	1+i	-1	0	1	-1-i	-i	1-i	2-i
i	1+i	-1	0	1	-1-i	-i	1-i	2-i	-1+i
1+i	-1	0	1	-1-i	-i	1-i	2-i	-1+i	i
-1	0	1	-1-i	-i	1-i	2-i	-1+i	i	1+i

Table S.3 Addition table for Exercise 14.4.3

$6\sqrt{-5}, 21\rangle = \langle 3 \rangle$. (b) $\langle 7, 1+2\sqrt{-5} \rangle \langle 7, 1-2\sqrt{-5} \rangle = \langle 49, 7-14\sqrt{-5}, 7+14\sqrt{-5}, 21 \rangle = \langle 7 \rangle$.
 (c) $\langle 3, 1+2\sqrt{-5} \rangle \langle 7, 1+2\sqrt{-5} \rangle = \langle 21, 7+14\sqrt{-5}, 3+6\sqrt{-5}, (1+2\sqrt{-5})^2 \rangle$. Since $21 = (1+2\sqrt{-5})(1-2\sqrt{-5})$, $1+2\sqrt{-5}$ divides all generators. $(7+14\sqrt{-5}) - 2(3+6\sqrt{-5}) = 1+2\sqrt{-5}$ implies $(1+2\sqrt{-5}) \in \langle 21, 7+14\sqrt{-5}, 3+6\sqrt{-5}, (1+2\sqrt{-5})^2 \rangle$. Thus, $\langle 21, 7+14\sqrt{-5}, 3+6\sqrt{-5}, (1+2\sqrt{-5})^2 \rangle = \langle 1+2\sqrt{-5} \rangle$. (d) $\langle 3, 1-2\sqrt{-5} \rangle \langle 7, 1-2\sqrt{-5} \rangle = \langle 21, 7-14\sqrt{-5}, 3-6\sqrt{-5}, (1-2\sqrt{-5})^2 \rangle$. Since $21 = (1+2\sqrt{-5})(1-2\sqrt{-5})$, $1-2\sqrt{-5}$ divides all generators; $(1-2\sqrt{-5}) \mid \langle 21, 7-14\sqrt{-5}, 3-6\sqrt{-5}, (1-2\sqrt{-5})^2 \rangle$. Since $7-14\sqrt{-5} - 2(3-6\sqrt{-5}) = 1-2\sqrt{-5}$, $\langle 21, 7-14\sqrt{-5}, 3-6\sqrt{-5}, (1-2\sqrt{-5})^2 \rangle \mid \langle 1-2\sqrt{-5} \rangle$. **14.2.3** (a) $\langle 6 \rangle \langle 3 \rangle = \langle 18 \rangle$. (b) $6 \notin \langle 18 \rangle$, so $6 \notin \langle 18 \rangle X$ for any ideal $X \subset \mathbb{Z}$. **14.2.5** Suppose I divides every ideal of R . In particular, $I \mid \langle 2 \rangle$, so that $2 \in I$. Also $I \mid \langle 3 \rangle$, so $3 \in I$. But $2, 3 \in I$ implies $3-2=1 \in I$, so $I = \langle 1 \rangle = R$. **14.3.1** Suppose R is a finite integral domain. Let x be a nonzero element of R and consider $xa = xb$. Then $xa - xb = x(a-b) = 0$. Since R is an integral domain $a-b=0$. Thus, for any distinct a, b , $ax \neq bx$. $|\langle x \rangle|$ is equal to the number of $r \in R$ such that $ax = r$ for some $a \in R$, which is the number of distinct elements a of R . Thus, $\langle x \rangle = R$ for any $x \in R$. R is a field. **14.4.3** One residue system is $\{0, 1, -1, i, -i, 1+i, 1-i, -1+i, -1-i, 2-i\}$; see Tables S.3 and S.4. The ring is isomorphic to \mathbb{Z}_{10} . It is not a field. **14.4.5** (a) Let $I = \{f \in F \mid f(0)=0\}$. If $f, g \in I$, then $(f+g)(0) = f(0)+g(0) = 0+0=0$ so $f+g \in I$. If $h \in F, f \in I$, then $(h \cdot f)(0) = h(0) \cdot f(0) = h(0) \cdot 0 = 0$ so $h \cdot f \in I$. Thus, I is an ideal. Suppose $0 \neq h \in F/I$. $h(0) = a \neq 0$. Then $f(x) = p(x) - a \in I$. $h(x) \equiv h(x) - f(x) \pmod{I}$, but $h(x) - f(x) = a$ for all $x \in [0, 1]$, which is an invertible function. Thus, h is a unit and F/I is a field. I is a maximal ideal. (b) $\{f \in F \mid f(1/4)=0\}$, $\{f \in F \mid f(1/2)=0\}$, $\{f \in F \mid f(3/4)=0\}$, $\{f \in F \mid f(1)=0\}$. (c) $I_3 = \{f \in F \mid f(0)=f(1/2)=f(1)=0\} \subsetneq I_2 = \{f \in F \mid f(0)=f(1)=0\} \subsetneq I_1 = \{f \in F \mid f(0)=0\}$.

\times	1	$-1-i$	$-i$	$1-i$	$2-i$	$-1+i$	i	$1+i$	-1
1	1	$-1-i$	$-i$	$1-i$	$2-i$	$-1+i$	i	$1+i$	-1
$-1-i$	$-1-i$	$1-i$	$-1+i$	$1+i$	0	$-1-i$	$1-i$	$-1+i$	$1+i$
$-i$	$-i$	$-1+i$	-1	$-1-i$	$2-i$	$1+i$	1	$1-i$	i
$1-i$	$1-i$	$1+i$	$-1-i$	$-1+i$	0	$1-i$	$1+i$	$-1-i$	$-1+i$
$2-i$	$2-i$	0	$2-i$	0	$2-i$	0	$2-i$	0	$2-i$
$-1+i$	$-1+i$	$-1-i$	$1+i$	$1-i$	0	$-1+i$	$-1-i$	$1+i$	$1-i$
i	i	$1-i$	1	$1+i$	$2-i$	$-1-i$	-1	$-1+i$	$-i$
$1+i$	$1+i$	$-1+i$	$1-i$	$-1-i$	0	$1+i$	$-1+i$	$1-i$	$-1-i$
-1	-1	$1+i$	i	$-1+i$	$2-i$	$1-i$	$-i$	$-1-i$	1

Table S.4 Multiplication table for Exercise 14.4.3

Index

- 15-puzzle, 171
- Abel, Niels, 6, 51
- abelian group, 195
- abstract group, 193
- al-Khwarizmi, 2
- Algebra*, 5
- algebraic expression, 23
- algebraic solution, 23
- algebraically resolvable, 24
- alternating group, 186
- Archimedes, 1
- argument, 11
- argument principle, 14
- Arithmetica*, 278
- Ars Magna*, 5
- ascending chain of ideals, 364
- associates, 295, 320
- automorphism, 205
- Binet's Formula, 90
- binomial coefficient, 76
- Binomial Theorem, 77, 101
- Bombelli, Rafael, 5
- Brahmagupta, 278
- cancelable ideal, 337
- Cardano, Gerolamo, 1
- Carmichael numbers, 89
- Cartesian number, 38
- Cartesian representation, 10
- Cauchy, Augustin-Louis, 6
- center, 214, 266
- centralizer, 214, 265
- Chinese Remainder Theorem, 69
- class equation, 267
- codomain, 422
- common divisor, 62
- commutative group, 195
- complement, 421
- complete residue system, 370
- complex integers, 356
- complex number, 9
- composite, 37
- composite ideal, 359
- composition of functions, 422
- congruent modulo n , 57
- conjugacy class, 234, 265
- conjugate, 14, 234, 265, 318
- conjugate ideal, 331
- conjunction, 413
- constant polynomial, 103
- constructible, 26
- contrapositive, 416
- converse, 416
- coset, 207
- cubic equation, 4
- cycle, 161
 - decomposition, 162
- cyclic group, 215
- cyclic permutation, 161
- cyclic table, 135
- cyclotomic equation, 50
- de Laplace, Pierre-Simon, 6
- decomposable group, 264
- decomposition
 - into disjoint cycles, 162
- Dedekind cuts, 314
- degree, 103
- del Ferro, Scipione, 1
- derivative, 154
- dihedral group, 188
- Diophantus, 278
- direct product, 261
- discriminant, 169
 - of cubic, 49
- disjoint cycle decomposition, 162
- disjunction, 413
- Disquisitiones Arithmeticae*, 26, 50
- distinct variants, 156
- divisible ideal, 359

- division of polynomials, 103
- divisor, 37
- divisors, 108
- domain, 422
- doubling of a cube, 34
- elementary symmetric polynomials, 121
- emptyset, 421
- equal ideals, 358
- equation
 - cubic, 4
 - first-degree, 2
 - quadratic, 3
- equivalence relation, 425
- equivalent propositions, 413
- Erlanger Programm, 187
- Euclidean algorithm, 63
- Euclidean domain, 362
- Euclidean function, 362
- Euclidean greatest common divisor, 115
- Euler φ -function, 93
- Euler's Criterion, 281
- Euler, Leonhard, 6, 33
- Eulerian integers, 304
- even permutation, 170
- existential quantifier, 418
- extension, 246
- factorable, 108
- factorization, 108
- factors, 108
- Fermat's Theorem, 86
- Ferrari, Lodovico, 49
- Fibonacci numbers, 84
- field, 99
 - extension, 246
 - isomorphism, 243
- field isomorphism, 243
- fifth-degree equation, 49
- first-degree equation, 2
- fourth-degree equation, 49
- Fundamental Theorem of Algebra, 51
- Fundamental Theorem of Symmetric Polynomials, 122
- Galois field, 135
- Galois imaginaries, 132
- Galois polynomial, 142
- Galois, Évariste, 50, 51
- Gauss, Carl Friedrich, 6, 26
- Gaussian integers, 294
- Gaussian primes, 294
- generator
 - of a group, 215
- greatest common divisor, 62, 113, 340
 - Euclidean, 115
- ground field, 102
- group, 193
 - abelian, 195
 - alternating, 186
 - center, 266
 - center of, 214
 - class equation, 267
 - commutative, 195
 - cyclic, 215
 - decomposable, 264
 - dihedral, 188
 - direct product, 261
 - generator of, 215
 - indecomposable, 264
 - isomorphism, 202
 - Klein 4-, 187
 - order, 202
 - Quaternion, 195
 - quotient, 230
 - symmetric, 184
- group isomorphism, 202
- group of permutations, 184
- highest common factor, 62
- Hilbert, David, 355
- Hisab al-jabr w'al-muqa-balah*, 3
- homomorphism, 234
- ideal, 358
 - cancelable, 337
 - composite, 359
 - divisible, 359
 - irreducible, 359
 - maximal, 368
 - principle, 358
 - unit, 358
- ideal multiplication, 335
- ideals, 322
- identity permutation, 158
- imaginary number, 10
- indecomposable integers, 313
- index, 210
- injective, 423
- integers, 356
 - complex, 356
 - irrational, 356
 - rational, 356

- integral domain, 361
- intersection, 421
- invariant function, 156
- irrational, 294
- irrational integers, 356
- irreducible, 108, 320, 356
- irreducible ideal, 359
- isomorphic, 370
 - fields, 242
- isomorphic groups, 202
- isomorphism, 202, 243
- Jordan, Camille, 207
- kernel, 237
- Khayyam, Omar, 1
- Klein 4-group, 187
- Klein, Felix, 187
- Kummer, Ernst, 310
- Lagrange's method, 127
- Lagrange's Theorem, 207
- Latin square, 196
- lattice point, 292
- Law of Homomorphisms, 367
- Law of Quadratic Reciprocity, 70, 292
- left inverse, 423
- Legendre symbol, 285
- Lindemann, Ferdinand, 34
- maximal ideal, 368
- method of false position, 2
- method of generating functions, 81
- method of infinite descent, 283
- minimal polynomial, 152
- modular arithmetic, 57
- modular order, 87
- modular roots of unity, 87
- modulus, 11
- monic polynomial, 103
- monomorphism, 235
- Multinomial Theorem, 91
- multiple, 37
- multiplicatively perfect, 73
- multiplicity of a zero, 110
- negation, 413
- Newton-Raphson method, 51
- norm, 295, 318
- normal subgroup, 228
- number
 - complex, 9
 - imaginary, 10
- odd permutation, 170
- On the Theory of Complex Numbers*, 310
- On the Theory of Numbers*, 131
- order, 136
 - modular, 87
 - of a finite group, 202
 - of a root of unity, 36
- parity of a permutation, 171
- Pascal's Identity, 77
- Pascal's Triangle, 77
- perfect number, 72
- permutation, 158, 425
 - cyclic, 161
 - even, 170
 - odd, 170
 - parity of, 171
- polar form, 12
- polynomial
 - irreducible, 108
 - minimal, 152
 - monic, 103
 - variants, 155
- polynomial over a field, 101
- prime, 37, 357
- prime factorization, 365
- prime ideal, 343
- prime subfield, 370
- primitive, 135, 150, 274
- Primitive Element Theorem, 144
- primitive root, 88
- principal ideal, 323
- principal ideal domain, 363
- principal ideals, 358
- product, 326
- propositional calculus, 413
- Pythagorean triple, 274
- quadratic domains, 356
- quadratic equation, 3
- quadratic field, 318
- quadratic nonresidue, 281
- quadratic residue, 281
- quartic equation, 49
- Quaternion group, 195
- quintic equation, 49
- quotient group, 230
- range, 422
- rational integers, 294, 356
- rational primes, 294
- reducible, 108
- relation, 425

- relatively prime, 63, 116
- Rhind Mathematical Papyrus*, 1
- right inverse, 424
- ring, 355
- ring isomorphism, 370
- roots of unity, 18
- RSA encryption, 95
- Ruffini, Paolo, 157
- signature, 236
- solvable by radicals, 24
- subfield, 246
- subgroup, 206
 - generator of, 215
 - index of, 210
 - normal, 228
 - proper, 207
 - trivial, 207
- subset, 421
- superset, 421
- surjective, 424
- symmetric group, 184
- Tartaglia, Niccolò, 1
- Taylor, Richard, 278
- trace, 318
- Traité des Substitutions*, 207
- transcendental numbers, 34
- transposition, 163
- trisection of an angle, 34
- trivial subgroup, 207
- truth tables, 413
- union, 421
- unique factorization property, 365
- unit, 320
- unit ideal, 358
- unit segment, 26
- units, 294
- universal quantifier, 418
- universe, 421
- variables, 101
- variants, 155
- vertex symmetries, 187
- Wantzel, Pierre, 35
- Wiles, Andrew, 278
- Wilson's Theorem, 280
- zero divisor, 361
- zero of a polynomial, 107
- zero polynomial, 103

Notation

A_n , 186	i , 9	$\sqrt[n]{z}$, 17
$\arg z$, 11	\mathfrak{a} , 323	$\mathfrak{o}(\zeta)$, 36, 136
$\begin{pmatrix} a \\ b \end{pmatrix}$, 76	1_G , 193	$\mathfrak{o}(a)$, 203
$Z(G)$, 214	Id , 158	$\mathfrak{o}(r)$, 87
Z_a , 214	$[G : H]$, 210	$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \end{pmatrix}$, 159
\mathbb{C} , 100	\mathbb{Z} , 23	$\varphi(n)$, 93
$C(a)$, 234	\mathbb{Z}_p^* , 195	$F[x, \leq n]$, 195
\bar{z} , 14	\mathbb{Z}_n , 58	$F[x]$, 102
$\bar{\mathfrak{a}}$, 331	$\mathbb{Z}[\sqrt{-5}]$, 72	G/H , 230
aH , 207	a^\dagger , 193	\mathbb{Q} , 100
$[M(x)]$, 244	$G \cong H$, 202	\mathbb{Q}^* , 294
$(a \ b \ c \ d)$, 161	$\text{Ker } f$, 237	\mathbb{R} , 100
D_n , 188	K , 192	$\langle \sigma \rangle$, 185
$G \times H$, 261	$\begin{pmatrix} a \\ p \end{pmatrix}$, 285	$S \cdot T$, 225
Δ_n , 169	$ z $, 11	S^{-1} , 225
$m \mid n$, 37	$N(z)$, 72	$\sum r_1 r_2 \cdots r_k$, 121
$a \equiv b \pmod{n}$, 57	$N(\alpha)$, 318	$\Delta(A, B)$, 358
$\asymp x \div$, 283	$N(a + bi)$, 295	S_n , 184
$\text{GF}(p, P(x))$, 135	$\ \sigma\ $, 165	$\text{Tr}(\alpha)$, 318
(m, n) , 62	$N(\mathfrak{a})$, 341	\mathbb{T} , 236